The Genetic Code as Physicochemical and Topological Geometry: Nucleotide-Determined Hydropathy, Exact Pseudoinverse Mapping, and the G-A-C-U Maximum

Douglas C. Youvan

doug@youvan.com

www.youvan.ai

June 14, 2026

The standard genetic code is usually presented as a table connecting sixty-four codons to twenty amino acids and translation stops, but this format conceals a coordinated physicochemical and topological organization. The original 1990 singular-value-decomposition analysis showed that amino-acid hydropathy can be predicted from twelve nucleotide-position variables, with uracil in the second codon position providing the dominant hydrophobic contribution. Later work placed this relationship in a broader mathematical framework: the complete 64 x 12 codon matrix has rank ten, admits an exact Moore–Penrose pseudoinverse, and projects amino-acid properties onto nucleotide-position coefficients through a transparent analytical map. Hydropathy and residue volume also segregate visibly according to the second codon position. A complementary 4 x 4 x 4 codon cube reveals further structure. When codons are collapsed to a 21-vertex amino-acid and stop multigraph, synonymous adjacencies become self-loops. Exhaustive enumeration of all twenty-four isotropic nucleotide orders shows that G-A-C-U and its mirror U-C-A-G uniquely maximize these loops, attaining forty-four. Together, these results connect amino-acid chemistry, low-dimensional codon algebra, Hamming geometry, and graph-theoretic compactness in one reproducible analysis of the standard genetic code itself.

Keywords: genetic code, nucleotide-determined hydropathy, Kyte–Doolittle hydropathy, amino-acid residue volume, physicochemical mapping, second codon position, uracil hydropathy determinant, 64 x 12 codon matrix, rank-10 codon algebra, Moore–Penrose pseudoinverse, exact pseudoinverse, Hamming geometry, codon cube, G-A-C-U nucleotide ordering, amino-acid multigraph, synonymous self-loops, graph theory, topological compactness, codon degeneracy, genetic-code topology. 129 pages.

OpenAI GPT-5.5 Thinking. CC4.0. https://orcid.org/0009-0004-4206-4253

Outline

1. From Genetic-Code Table to Physicochemical and Topological Map

1.1 The conventional genetic-code table

Introduce the sixty-four codons, twenty amino acids, translation stops, and synonymous degeneracy.

1.2 Structure concealed by the tabular representation

Explain how the conventional table obscures positional effects, physicochemical clustering, mutation geometry, and synonymous connectivity.

1.3 The two principal questions

Ask how strongly amino-acid properties map onto nucleotide positions and whether a particular nucleotide ordering produces an extremal topology of synonymous codons.

1.4 Scope of the paper

State that the analysis concerns measurable algebraic, physicochemical, geometric, and graph-theoretic properties of the standard genetic code.

2. Nucleotide-Determined Hydropathy in the 1990 SVD Analysis

2.1 Origins in membrane-protein engineering

Describe the practical motivation for predicting encoded amino-acid hydropathy directly from nucleotide sequence.

2.2 The 61 x 12 sense-codon matrix

Define the twelve one-hot nucleotide-position variables for the sixty-one sense codons.

2.3 The SVD solution

Present the minimum-norm least-squares mapping from sense-codon hydropathy values to twelve nucleotide-position coefficients.

2.4 The twelve hydropathy coefficients

Display the coefficient matrix for A, G, C, and U at codon positions one, two, and three.

2.5 The dominant second-position effect

Emphasize that U in the second codon position supplies the largest hydrophobic contribution and A supplies the strongest hydrophilic contribution.

2.6 Codon-level hydropathy reconstruction

Show how the three nucleotide-position contributions are added to predict the hydropathy associated with a codon.

3. Hydropathy and Residue-Volume Segregation

3.1 Two-dimensional amino-acid property space

Define Kyte–Doolittle hydropathy and amino-acid residue volume as complementary physicochemical coordinates.

3.2 Classification by the second codon position

Organize the code into the NUN, NCN, NAN, and NGN families.

3.3 Physicochemical interpretation of the four families

Describe NUN as predominantly hydrophobic, NCN as relatively small, NAN as predominantly hydrophilic, and NGN as more heterogeneous.

3.4 Stop-codon padding for complete-cube analysis

Assign zero hydropathy to X and the mean amino-acid residue volume to X.

3.5 Figure 1 placement: Hydropathy–volume segregation

Place Figure 1 immediately after the four second-position families and padding conventions are explained.

Caption: Amino-acid segregation in Kyte–Doolittle hydropathy and residue-volume space according to the second nucleotide of the codon. Stop X is assigned zero hydropathy and the mean amino-acid residue volume for the complete-codon calculation.

3.6 Quantitative interpretation

Relate the visible segregation to the nucleotide-position coefficients and the dominant role of the middle nucleotide.

4. Complete Codon Space, Rank Ten, and the Exact Pseudoinverse

4.1 The complete 64 x 12 codon matrix

Construct the one-hot matrix containing every possible four-letter triplet.

4.2 Complete factorial balance

Show that every nucleotide occurs sixteen times at each position and every ordered pair occurs four times across two positions.

4.3 Rank-ten structure

Derive the two positional-block dependencies and the resulting rank of ten.

4.4 The exact Moore–Penrose pseudoinverse

Present the closed analytical formula for the complete codon matrix and its two possible element values.

4.5 Brief relationship to the historical SVD analysis

Explain in one concise subsection that the sixty-one-codon SVD reproduces the original biological coefficients, whereas the padded sixty-four-codon formula preserves complete symmetry. Avoid repeating this distinction elsewhere.

5. Mapping Amino-Acid Properties onto Nucleotide Positions

5.1 The codon-to-amino-acid assignment matrix

Define the matrix that transfers amino-acid property values to the codons that encode them.

5.2 Amino-acid property vectors and matrices

Represent hydropathy, residue volume, and additional properties in a unified matrix.

5.3 From amino acids to codon property fields

Construct the hydropathy and volume values assigned across codon space.

5.4 From codon fields to twelve nucleotide coefficients

Apply the pseudoinverse to map codon-level properties onto nucleotide identity and codon position.

5.5 Reconstruction from nucleotide contributions

Show how the nucleotide coefficients reconstruct the additive component of each codon property.

5.6 Biological interpretation

Interpret the coefficients as quantitative measures of the relationship between codon composition and encoded amino-acid chemistry.

6. Hamming Geometry of the Codon Projection

6.1 The projection operator

Define the orthogonal projection from the sixty-four-dimensional codon property space onto the ten-dimensional additive nucleotide-position space.

6.2 Matching positions and Hamming distance

Define positional agreement and the number of nucleotide substitutions separating two codons.

6.3 The exact Hamming-distance kernel

Derive the projection coefficient as a function only of codon Hamming distance.

6.4 Recoverable and residual property structure

Separate hydropathy or volume into an additive nucleotide-recoverable component and a residual interaction component.

6.5 Hamming graph versus ordered cube

Distinguish the 288-edge full single-substitution Hamming graph from the 144-edge ordered codon cube used in the topology analysis.

7. The Isotropic G-A-C-U Codon Cube

7.1 The 4 x 4 x 4 codon lattice

Represent the three codon positions as three coordinate axes with four nucleotide states per axis.

7.2 Isotropic nucleotide ordering

Define an isotropic cube as one using the same nucleotide order on all three axes.

7.3 The twenty-four isotropic permutations

Describe the complete set of possible common nucleotide orders.

7.4 The G-A-C-U coordinate system

Define the consecutive axis path G–A–C–U and its reversed mirror U–C–A–G.

7.5 Figure 2 placement: The G-A-C-U codon cube

Place Figure 2 immediately after defining the G-A-C-U coordinate system and before introducing the collapsed multigraph.

Caption: The 4 x 4 x 4 codon cube with the first, second, and third codon-position axes all ordered G-A-C-U. Each lattice site represents one codon and is labeled by its encoded amino acid or stop X.

7.6 Relationship between the cube and codon matrix

Explain that the matrix is order independent, whereas the cube introduces an order-dependent nearest-neighbor topology.

8. Collapse to the 21-Vertex Amino-Acid Multigraph

8.1 The 144 nearest-neighbor lattice edges

Derive the number of edges in the ordered 4 x 4 x 4 cube.

8.2 Collapse by translated identity

Map the sixty-four codon vertices onto twenty amino-acid vertices and stop X while retaining every lattice edge.

8.3 Parallel edges

Explain why multiple codon-level adjacencies can connect the same pair of amino-acid vertices.

8.4 Synonymous adjacencies as self-loops

Define a self-loop as an ordered-cube edge whose two codons encode the same translated label.

8.5 A Word-friendly loop-count definition

First define the synonymous-edge indicator, then define the total loop count using a short summation expression.

8.6 Figure 3 placement: The G-A-C-U amino-acid multigraph

Place Figure 3 immediately after the self-loop definition.

Caption: The G-A-C-U amino-acid multigraph obtained by collapsing sixty-four codons to twenty amino-acid vertices and stop X while retaining all 144 lattice connections. The forty-four synonymous adjacencies appear as self-loops.

9. Exhaustive Enumeration of Isotropic Nucleotide Orders

9.1 Computational procedure

Describe the independent construction of all twenty-four isotropic cubes and the counting of synonymous nearest-neighbor edges.

9.2 The maximum loop count

Establish that the greatest observed count is forty-four.

9.3 The maximizing orders

Show that only G-A-C-U and its reversed mirror U-C-A-G attain the maximum.

9.4 Uniqueness up to reversal

Treat the two oriented orders as one undirected geometric solution.

9.5 Comparison with the next-highest orders

Report the loop counts immediately below the maximum to show that the result is an actual extremum rather than a broad tie.

9.6 Reproducibility

Describe the Python implementation, assertions, edge list, and full permutation table required to reproduce the result.

10. Biological and Topological Interpretation of the Forty-Four Loops

10.1 Distribution of loops among translated labels

Present the contributions of leucine, arginine, serine, the fourfold-degenerate amino acids, and stop X.

10.2 Third-position degeneracy

Explain why many loops occur along the third codon-position axis.

10.3 Sixfold-degenerate amino acids

Examine how leucine, arginine, and serine contribute through more than one codon block.

10.4 Synonymous compactness

Interpret the loop count as a transparent measure of local clustering of codons with identical translated output.

10.5 Relationship to codon degeneracy

Distinguish the number of codons assigned to an amino acid from the spatial compactness of those codons under a selected nucleotide order.

10.6 Limits of the loop statistic

Clarify that maximum self-loop count is one defined topological criterion and does not imply optimization under every possible graph measure.

11. Controls, Alternative Graph Measures, and Extensions

11.1 Separation of universal and code-specific results

Distinguish rank, pseudoinverse, and Hamming geometry from physicochemical alignment and the G-A-C-U topology.

11.2 Randomized-code controls

Compare the observed property projections and maximum loop count with unrestricted, degeneracy-preserving, and synonymous-block-preserving randomized codes.

11.3 Alternative physicochemical scales

Test whether the second-position hydropathy and residue-volume patterns persist across established property scales.

11.4 Alternative topological measures

Evaluate connector length, synonymous component number, pairwise distance, graph smoothness, and physicochemical edge cost.

11.5 Anisotropic axis orderings

Extend the search from twenty-four isotropic orders to all independently ordered axis triples.

11.6 Statistical reporting

Specify effect sizes, permutation probabilities, confidence intervals, and preregistered comparison metrics.

12. Synthesis and Conclusions

12.1 Nucleotide-determined amino-acid chemistry

Summarize the strong mapping of hydropathy and residue volume onto codon-position coordinates.

12.2 Low-dimensional codon algebra

Summarize the rank-ten structure, exact pseudoinverse, and Hamming-distance projection without revisiting the historical distinction at length.

12.3 Codon-cube topology

Summarize the construction of the isotropic cube and its collapse to a 21-vertex multigraph.

12.4 The G-A-C-U maximum

State that G-A-C-U and U-C-A-G uniquely attain forty-four synonymous self-loops among all isotropic orders.

12.5 Unified interpretation

Conclude that the standard genetic code combines physicochemical mapping, low-dimensional algebra, symbolic geometry, and extremal synonymous topology.

12.6 Final conclusion

State the demonstrated mathematical results directly, without discussion of genetic-code origins.

13. References

Collect the original 1990 SVD study, amino-acid property scales, exact pseudoinverse derivations, codon-cube and graph analyses, computational methods, randomized-code literature, and figure sources here. Sections 1–12 will contain no inline references, hyperlinks, or video links.

Postscript

1. From Genetic-Code Table to Physicochemical and Topological Map

1.1 The Conventional Genetic-Code Table

The standard genetic code is commonly presented as a table that assigns sixty-four nucleotide triplets to twenty amino acids and translation-stop signals. Each codon contains three ordered nucleotide positions, and each position can contain adenine, cytosine, guanine, or uracil. The total number of possible codons is therefore

$$4^3 = 64.$$

Most amino acids are encoded by more than one codon. This redundancy, usually called degeneracy, is distributed unevenly. Methionine and tryptophan each have one codon, several amino acids have two or four codons, isoleucine has three, and leucine, arginine, and serine each have six. Three codons ordinarily function as translation stops and are represented here by X.

The table is effective for its primary purpose: translating nucleotide sequence into amino-acid sequence. Given a codon c, the table returns its translated label $\phi(c)$. Formally, the genetic code can be written as a mapping

$$\phi: \mathcal{C} \to \mathcal{R},$$

where $\mathcal{C}$ is the set of sixty-four codons and $\mathcal{R}$ contains the twenty amino acids and stop X.

In this representation, the code appears primarily as a collection of assignments. UUU encodes phenylalanine, AUG encodes methionine, GCU encodes alanine, and UGG encodes tryptophan. This assignment view is biologically indispensable, but it does not reveal the full mathematical organization of the code.

The printed arrangement of a codon table is also partly conventional. Rows and columns are chosen for readability, not necessarily to preserve every natural relationship among codons. Two codons that differ by one nucleotide may appear far apart in the table, while visually adjacent entries may not correspond to equivalent mutational or physicochemical relationships.

The tabular format therefore answers the question

Which amino acid does this codon encode?

but it is less effective for addressing questions such as

How are amino-acid properties distributed across codon space?

or

How compactly are synonymous codons connected?

These questions require the code to be represented not only as a lookup table, but also as a numerical matrix, a geometric lattice, and a graph.

1.2 Structure Concealed by the Tabular Representation

The first major structure hidden by the conventional table is the unequal contribution of the three codon positions. The first, second, and third nucleotides do not play equivalent roles in determining the physicochemical identity of the encoded amino acid.

The third position is often associated with synonymous variation. In many codon families, changes at the third nucleotide leave the encoded amino acid unchanged. The first position commonly refines amino-acid identity within broader codon classes. The second position has an especially strong relationship to amino-acid hydropathy and residue volume.

If the second nucleotide is fixed while the first and third positions are allowed to vary, the code separates into four families:

$$NUN, NCN, NAN, NGN,$$

where N denotes any nucleotide.

These families occupy recognizable regions of amino-acid physicochemical space. The NUN family is dominated by hydrophobic residues. The NCN family contains relatively small residues. The NAN family is enriched in hydrophilic, polar, and charged residues. The NGNfamily is more heterogeneous, spanning a broad range of residue volumes and hydropathic values.

This organization is visible when amino acids are plotted using hydropathy and residue volume as coordinates. It is also measurable through a nucleotide-position model in which each codon is represented by twelve binary variables: four possible nucleotides at each of three positions.

The complete codon matrix has the form

$$D \in \{0,1\}^{64\times 12}.$$

For codon c, position p, and nucleotide s,

$$D_{c,(p,s)} = \begin{cases} 1, & c_p = s, \\ 0, & c_p \neq s. \end{cases}$$

Each row contains exactly three ones, one within each positional block.

This matrix representation makes it possible to ask how strongly an amino-acid property field aligns with independent nucleotide-position effects. Hydropathy, residue volume, or another numerical property can be assigned to every codon through the amino acid it encodes. The resulting codon property field can then be projected onto the twelve nucleotide-position variables.

A second hidden structure is geometric. The sixty-four codons can be arranged on a $4 \times 4 \times 4$ lattice, with the three axes corresponding to the three codon positions. Each lattice site represents one codon.

A third hidden structure is topological. Once an order is chosen for the four nucleotides along each axis, neighboring lattice sites define codon-codon connections. If two neighboring codons encode the same amino acid, their connection is synonymous. When all codons with the same translated label are collapsed to one vertex, synonymous connections become self-loops.

The genetic code therefore supports several complementary descriptions:

assignment structure,
physicochemical structure,
matrix structure,
geometric structure,

and

graph-theoretic structure.

The purpose of the present analysis is to connect these descriptions rather than treating them as unrelated observations.

1.3 The Two Principal Questions

The paper is organized around two principal questions.

The first is physicochemical:

How strongly do amino-acid properties map onto nucleotide positions?

The original hydropathy analysis addressed this question using the sixty-one sense codons. Each codon was represented by twelve nucleotide-position indicators, and singular-value decomposition was used to obtain twelve coefficients. These coefficients measured the contribution of each nucleotide at each position to the hydropathy of the encoded amino acid.

The strongest result was the large positive contribution of uracil in the second codon position:

$$\beta_{2,U} \approx 4.02.$$

Adenine in the second position supplied the strongest negative contribution:

$$\beta_{2,A} \approx -2.92.$$

Thus, the middle nucleotide emerged as the dominant determinant of nucleotide-predicted hydropathy.

The same framework can be extended to residue volume and other amino-acid properties. In general, if Q contains amino-acid property values and Gmaps those values onto codons, then the codon property field is

$$Y = GQ.$$

The corresponding nucleotide-position coefficients are obtained from

$$B = D^{+}GQ.$$

This equation expresses the physicochemical question mathematically. It asks how amino-acid chemistry is represented in the coordinate system of codon nucleotides and positions.

The second principal question is topological:

Which isotropic nucleotide order produces the greatest synonymous compactness?

An isotropic codon cube uses the same nucleotide order along all three codon-position axes. Since there are four nucleotides, there are

$$4! = 24$$

possible isotropic orders.

For each order π, the $4 \times 4 \times 4$ lattice contains

$$3(4-1)4^2 = 144$$

nearest-neighbor edges.

The codons are then collapsed to twenty amino-acid vertices and stop X. A nearest-neighbor edge becomes a self-loop when its two codons have the same translated label.

To make the definition easy to read and transfer, define the synonymous-edge indicator by

$$I_\phi(e) = \begin{cases} 1, & \text{the endpoints of } e \text{ have the same label,} \\ 0, & \text{otherwise.} \end{cases}$$

The total self-loop count for nucleotide order $\pi$is

$$L(\pi) = \sum_{e \in E_\pi} I_\phi\,(e).$$

Exhaustive enumeration shows that

$$\max_{\pi \in S_4} L(\pi) = 44,$$

and the maximizing orders are

$$GACU$$

and its reversed mirror

$$UCAG.$$

The physicochemical and topological questions are related but distinct. The first examines how amino-acid properties align with nucleotide-position variables. The second examines how synonymous codons cluster under an ordered geometric embedding.

Together, they test whether the genetic code contains structure beyond the individual codon-to-amino-acid assignments listed in the conventional table.

1.4 Scope of the Paper

This paper develops a mathematical analysis of the standard genetic code itself. Its scope is limited to measurable and reproducible relationships among codon composition, amino-acid physicochemical properties, codon geometry, and graph topology.

The physicochemical analysis begins with the historical result that amino-acid hydropathy can be predicted from twelve nucleotide-position variables. The discussion then extends to residue volume and to the visible segregation of amino acids according to the second codon position.

The algebraic analysis introduces the complete 64×12 codon matrix, its rank-ten structure, and its exact Moore–Penrose pseudoinverse. The distinction between the original sixty-one-codon SVD calculation and the padded sixty-four-codon exact formulation will be stated briefly and then set aside. Both formulations address the same central mapping from codon-level properties to nucleotide-position coefficients.

The geometric analysis represents the sixty-four codons as a $4 \times 4 \times 4$ lattice. The Hamming-distance projection describes the order-independent geometry of complete codon space, while the ordered cube introduces an explicit nucleotide path along each axis.

The graph-theoretic analysis collapses the sixty-four codons to twenty amino-acid vertices and stop X, retaining all 144 ordered-cube edges. Synonymous adjacencies become self-loops, producing a transparent measure of local codon compactness.

The paper will establish four principal results:

amino-acid hydropathy maps strongly onto nucleotide positions,
the complete codon matrix has rank ten and an exact pseudoinverse,
the projection kernel depends only on codon Hamming distance,

and

$GACU$ and $UCAG$ uniquely maximize isotropic self-loops at 44.

The objective is not to propose a theory of genetic-code origins. It is to describe and connect the mathematical structures that can be directly calculated from the standard code.

The resulting framework treats the genetic code as both a physicochemical map and a topological object. The following sections develop these two aspects in parallel and show how they converge in a unified description of codon organization.

2. Nucleotide-Determined Hydropathy in the 1990 SVD Analysis

2.1 Origins in Membrane-Protein Engineering

The original analysis arose from a practical problem in membrane-protein engineering. Hydropathy strongly influences whether an amino-acid side chain is exposed to water, buried within a protein, or positioned within a lipid bilayer. Transmembrane segments are therefore characterized by sustained runs of relatively hydrophobic residues, while loops and solvent-exposed regions generally contain more polar or charged amino acids.

The conventional procedure is to translate a nucleotide sequence into an amino-acid sequence and then calculate a hydropathy profile. The 1990 analysis asked whether part of this relationship could be detected directly at the nucleotide level. Specifically, it tested whether the identities and positions of the three nucleotides in a codon could predict the hydropathy of the encoded amino acid.

The proposed mapping was

$$\text{codon nucleotides} \longrightarrow \text{amino-acid hydropathy.}$$

The aim was not to claim that codon composition determines hydropathy exactly. Rather, the question was whether a simple additive model based on twelve nucleotide-position variables could recover a substantial and biologically interpretable fraction of the hydropathy signal.

This was a stringent test because the model used only nucleotide identity and codon position. It did not use amino-acid names, side-chain structures, or protein-level context.

2.2 The 61×12 Sense-Codon Matrix

The original calculation used the sixty-one sense codons and excluded the three stop codons. This avoided assigning artificial hydropathy values to translation termination signals.

Let

$$D_{61} \in \{0,1\}^{61\times 12}$$

be the sense-codon matrix.

The twelve columns represent four nucleotides at each of three codon positions. Using the order A, G, C, U, the columns are

$$(1, A), (1, G), (1, C), (1, U),$$
$$(2, A), (2, G), (2, C), (2, U),$$

and

$$(3, A), (3, G), (3, C), (3, U).$$

For codon c, position p, and nucleotide s,

$$\Big(D_{61}\big)_{c,(p,s)} = \begin{cases} 1, & c_p = s, \\ 0, & c_p \neq s. \end{cases}$$

Each row contains exactly three entries equal to one.

Let

$$\mathbf{h}_{61} \in \mathbb{R}^{61}$$

contain the Kyte–Doolittle hydropathy value of the amino acid encoded by each sense codon.

Synonymous codons receive the same hydropathy value because they encode the same amino acid.

The linear model is

$$D_{61}\boldsymbol{\beta} \approx \mathbf{h}_{61},$$

where

$$\boldsymbol{\beta} \in \mathbb{R}^{12}$$

contains the nucleotide-position hydropathy coefficients.

2.3 The SVD Solution

The matrix D_{61} is rectangular and rank deficient, so an ordinary matrix inverse cannot be used. Singular-value decomposition provides the minimum-norm least-squares solution.

Write

$$D_{61} = U\Sigma V^T.$$

The Moore–Penrose pseudoinverse is

$$D_{61}^{+} = V\Sigma^{+}U^{T}.$$

The nucleotide-determined hydropathy coefficients are then

$$\boxed{\boldsymbol{\beta}_{61} = D_{61}^{+}\mathbf{h}_{61}.}$$

The predicted hydropathy vector is

$$\boxed{\hat{\mathbf{h}}_{61} = D_{61}\boldsymbol{\beta}_{61}.}$$

Equivalently,

$$\boxed{\hat{\mathbf{h}}_{61} = D_{61}D_{61}^{+}\mathbf{h}_{61}.}$$

The analysis produced a correlation of approximately

$$r \approx 0.92$$

between observed amino-acid hydropathy and hydropathy predicted from the twelve nucleotide-position variables.

This was the principal biological result: codon composition alone retained a strong signal of encoded amino-acid hydropathy.

2.4 The Twelve Hydropathy Coefficients

The twelve coefficients can be displayed as a 3×4 matrix. Rows correspond to codon positions one, two, and three. Columns are ordered A, G, C, U.

$$\boxed{\boldsymbol{\beta}_{61} \approx \begin{pmatrix} -0.57 & 0.76 & -1.18 & 0.52 \\ -2.92 & -1.48 & -0.09 & 4.02 \\ -0.35 & -0.50 & 0.19 & 0.19 \end{pmatrix}.}$$

The first-position coefficients show moderate hydropathy differences.

The third-position coefficients are comparatively small, consistent with the frequent occurrence of synonymous variation at the third nucleotide.

The second-position coefficients are much larger in magnitude and span a wider range.

Their approximate values are

$$\beta_{2,A} = -2.92,$$
$$\beta_{2,G} = -1.48,$$
$$\beta_{2,C} = -0.09,$$

and

$$\beta_{2,U} = 4.02.$$

The range of the second-position coefficients is therefore

$$4.02 - (-2.92) = 6.94.$$

This is substantially larger than the corresponding ranges at positions one and three.

2.5 The Dominant Second-Position Effect

The strongest coefficient in the entire matrix is

$$\boxed{\beta_{2,U} \approx 4.02.}$$

This coefficient indicates that uracil in the second codon position makes the largest positive contribution to predicted hydropathy.

The result is consistent with the composition of the NUNcodon family. Middle-position uracil is associated with phenylalanine, leucine, isoleucine, methionine, and valine, all of which are hydrophobic.

The strongest negative coefficient is

$$\boxed{\beta_{2,A} \approx -2.92.}$$

Middle-position adenine is associated with several polar, charged, or hydrophilic amino acids, including aspartate, glutamate, asparagine, glutamine, lysine, and histidine.

The second-position pattern can therefore be summarized as

$$U \text{ at position } 2 \longrightarrow \text{hydrophobic contribution},$$

while

$$A \text{ at position } 2 \longrightarrow \text{hydrophilic contribution.}$$

Cytosine in the second position is nearly neutral in the additive hydropathy model, while guanine contributes moderately toward hydrophilicity.

The middle nucleotide does not uniquely determine the amino acid, but it strongly constrains the physicochemical class of the encoded residue.

2.6 Codon-Level Hydropathy Reconstruction

For codon

$$c = (c_1, c_2, c_3),$$

the predicted hydropathy is obtained by adding the three relevant nucleotide-position coefficients:

$$\boxed{\widehat{H}(c) = \beta_{1,c_1} + \beta_{2,c_2} + \beta_{3,c_3}.}$$

For example, methionine is encoded by

$$AUG.$$

Using the coefficient table,

$$\widehat{H}(AUG) = \beta_{1,A} + \beta_{2,U} + \beta_{3,G}.$$

Substitution gives

$$\widehat{H}(AUG) \approx -0.57 + 4.02 - 0.50.$$

Therefore,

$$\widehat{H}(AUG) \approx 2.95.$$

Methionine has positive Kyte–Doolittle hydropathy, so the model correctly places it in the hydrophobic region.

The residual for codon $c$is

$$r(c) = H(c) - \widehat{H}(c).$$

Across all sense codons, the residual vector is

$$\mathbf{r}_{61} = \mathbf{h}_{61} - \hat{\mathbf{h}}_{61}.$$

The residual contains the hydropathy structure not captured by independent additive nucleotide-position effects.

The significance of the 1990 result is therefore not exact prediction of every amino acid. It is the discovery that a simple twelve-variable codon model recovers a strong, chemically interpretable hydropathy signal.

The three central conclusions are:

hydropathy is substantially predictable from codon composition,
the second codon position is the dominant positional determinant,

and

second-position U supplies the largest hydrophobic contribution.

These findings provide the physicochemical foundation for the geometric and graph-theoretic analyses developed later in the paper.

3. Hydropathy and Residue-Volume Segregation

3.1 Two-Dimensional Amino-Acid Property Space

Hydropathy and residue volume describe two complementary aspects of amino-acid chemistry. Hydropathy measures the preference of an amino-acid side chain for a nonpolar environment rather than water. Residue volume measures the approximate space occupied by that residue within a folded protein.

Let

$$H_r$$

denote the Kyte–Doolittle hydropathy of amino acid r, and let

$$V_r$$

denote its residue volume.

Each amino acid can then be represented as a point in a two-dimensional physicochemical space:

$$\mathbf{q}_r = \begin{pmatrix} H_r \\ V_r \end{pmatrix}.$$

The horizontal axis represents hydropathy. Positive values correspond to relatively hydrophobic residues, while negative values correspond to relatively hydrophilic residues.

The vertical axis represents residue volume. Small residues appear toward the bottom of the plot, while larger residues appear toward the top.

These two properties are biologically related but not redundant. Hydropathy influences membrane insertion, solvent exposure, and formation of hydrophobic protein interiors. Residue volume influences steric packing, active-site geometry, conformational freedom, and the stability of folded protein structures.

For example, glycine has a very small residue volume, whereas tryptophan has a very large volume. Isoleucine, valine, leucine, and phenylalanine are strongly hydrophobic. Aspartate, glutamate, lysine, and arginine are strongly hydrophilic.

A two-dimensional hydropathy–volume plot therefore provides a compact representation of major amino-acid physicochemical differences.

The important genetic-code observation is that the amino acids do not occupy this plane independently of their codons. Their distribution is strongly related to the nucleotide present at the second codon position.

3.2 Classification by the Second Codon Position

Let

$$c = (c_1, c_2, c_3)$$

be a codon. The middle nucleotide is

$$c_2.$$

Fixing the second nucleotide while allowing the first and third positions to vary produces four codon families:

$$NUN,$$
$$NCN,$$
$$NAN,$$

and

$$NGN.$$

Here, Ndenotes any nucleotide.

Each family contains sixteen codons before synonymous codons are collapsed:

$$4 \times 4 = 16.$$

The four codon subsets may be written as

$$\mathcal{C}_U = \{c : c_2 = U\},$$
$$\mathcal{C}_C = \{c : c_2 = C\},$$
$$\mathcal{C}_A = \{c : c_2 = A\},$$

and

$$\mathcal{C}_G = \{c : c_2 = G\}.$$

The corresponding amino-acid groups are determined by the standard genetic-code assignment.

The NUN family includes phenylalanine, leucine, isoleucine, methionine, and valine. These residues are predominantly hydrophobic.

The NCN family includes serine, proline, threonine, and alanine. These residues are relatively small and occupy a compact lower region of the volume axis.

The NAN family includes tyrosine, histidine, glutamine, asparagine, lysine, aspartate, glutamate, and stop X. Most of the amino acids in this group are polar, charged, or hydrophilic.

The NGN family includes cysteine, tryptophan, arginine, serine, glycine, and stop X. This group is more heterogeneous than the others.

Serine belongs to two second-position families because it is encoded by both UCN and AGY-type codons. Stop X also appears in both the NAN and NGN groups.

These overlaps do not eliminate the broad physicochemical segregation. They reflect the actual structure and irregularity of the standard genetic code.

3.3 Physicochemical Interpretation of the Four Families

The most conspicuous separation occurs along the hydropathy axis.

The principal amino acids encoded by NUNcodons have strongly positive Kyte–Doolittle values:

$$H_I = 4.5,$$
$$H_V = 4.2,$$
$$H_L = 3.8,$$
$$H_F = 2.8,$$

and

$$H_M = 1.9.$$

This concentration of hydrophobic residues explains why uracil in the second codon position receives the largest positive coefficient in the nucleotide-determined hydropathy analysis:

$$\beta_{2,U} \approx 4.02.$$

The NUN family can therefore be summarized as

$$NUN \longrightarrow \text{predominantly hydrophobic amino acids.}$$

The NANfamily occupies the opposite side of the hydropathy axis. It includes several strongly hydrophilic amino acids:

$$H_K = -3.9,$$
$$H_D = -3.5,$$
$$H_E = -3.5,$$
$$H_N = -3.5,$$

and

$$H_Q = -3.5.$$

This organization explains the large negative coefficient associated with adenine in the second codon position:

$$\beta_{2,A} \approx -2.92.$$

Thus,

$$NAN \longrightarrow \text{predominantly hydrophilic amino acids.}$$

The NCN family is distinguished most clearly by residue size. Its principal residues have relatively small volumes:

$$V_A = 88.6\ \text{Å}^3,$$
$$V_S = 89.0\ \text{Å}^3,$$
$$V_P = 112.7\ \text{Å}^3,$$

and

$$V_T = 116.1\ \text{Å}^3.$$

Therefore,

$$NCN \longrightarrow \text{relatively small residues.}$$

The NGN family is more broadly distributed. It contains glycine, the smallest amino acid, and tryptophan, the largest. It also contains strongly hydrophilic arginine and moderately hydrophobic cysteine.

Therefore,

$$NGN \longrightarrow \text{heterogeneous physicochemical region.}$$

The second codon nucleotide can consequently be regarded as a coarse physicochemical classifier:

$$\boxed{c_2 \longrightarrow \text{broad amino-acid property class.}}$$

It does not uniquely identify the amino acid, but it substantially restricts the region of hydropathy–volume space in which that amino acid lies.

3.4 Stop-Codon Padding for Complete-Cube Analysis

The original sixty-one-codon SVD analysis excluded stop codons because stop signals do not encode amino-acid residues.

The complete 64×12 codon matrix includes all sixty-four triplets. To analyze a complete property field, values must therefore be assigned to the stop symbol X.

For hydropathy, the chosen neutral padding value is

$$\boxed{H_X = 0.}$$

This does not mean that a stop codon possesses a physical hydropathy of zero. It is a numerical convention that places X between the hydrophobic and hydrophilic extremes.

For residue volume, zero would place X far outside the range of ordinary amino-acid values. The chosen padding is therefore the mean volume of the twenty amino acids:

$$\boxed{V_X = \frac{1}{20}\sum_{r=1}^{20} V_r\,.}$$

For the residue-volume scale used here,

$$\boxed{V_X \approx 141.26\ \text{Å}^3.}$$

The complete hydropathy and volume vectors are then

$$\mathbf{h}_{64}$$

and

$$\mathbf{v}_{64}.$$

The padded values preserve the complete $4 \times 4 \times 4$ codon space and allow the exact pseudoinverse to be applied without deleting rows from the codon matrix.

The padding convention must be stated explicitly because it affects the numerical coefficients slightly. It does not alter the major qualitative result: second-position uracil remains the dominant positive hydropathy determinant.

3.5 Figure 1: Hydropathy–Volume Segregation

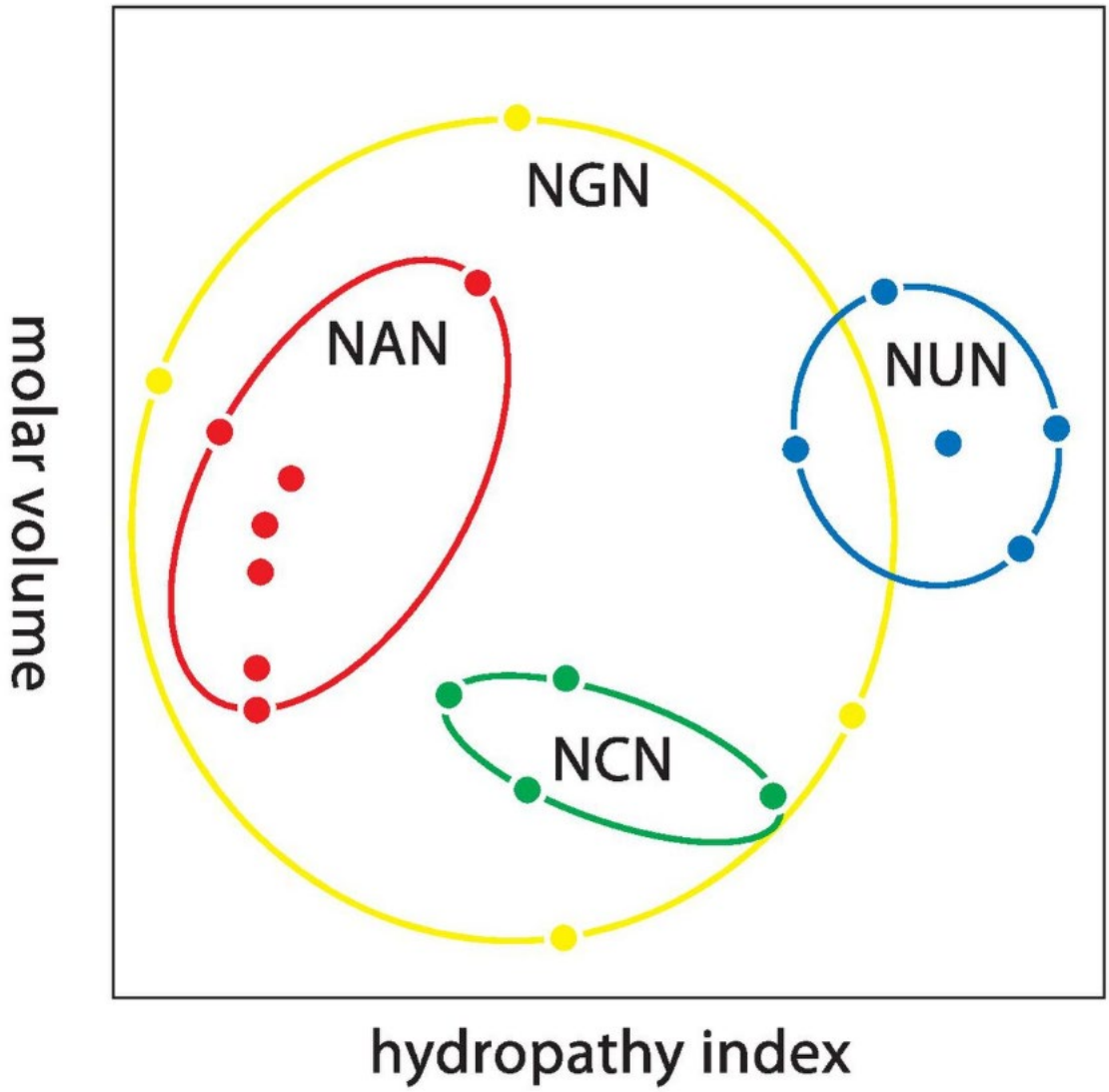

Figure 1. Hydropathy–Residue-Volume Segregation The NUN group occupies the hydrophobic right side. The NAN group occupies the hydrophilic left side. The NCN group is concentrated at relatively low residue volumes. The NGN group spans a wider range of both variables.

3.6 Quantitative Interpretation

The visible segregation can be quantified by mapping hydropathy and residue volume onto the twelve nucleotide-position variables.

Let

$$Q = (\mathbf{h} \quad \mathbf{v})$$

contain the two amino-acid property vectors.

Let G be the codon-to-amino-acid assignment matrix. The codon-level property matrix is

$$\boxed{Y = GQ.}$$

The nucleotide-position coefficients are

$$\boxed{B = D^{+}Y.}$$

Equivalently,

$$\boxed{B = D^{+}GQ.}$$

The reconstructed property matrix is

$$\boxed{\hat{Y} = DB.}$$

Thus,

$$\boxed{\hat{Y} = DD^{+}GQ.}$$

The first column of B contains hydropathy coefficients, while the second contains residue-volume coefficients.

For the complete padded hydropathy field, the observed-versus-predicted correlation is approximately

$$r_H \approx 0.904.$$

The corresponding explained variance is

$$R_H^2 \approx 0.817.$$

For residue volume, the correlation is approximately

$$r_V \approx 0.772,$$

with

$$R_V^2 \approx 0.596.$$

These results show that the additive nucleotide-position model captures a substantial fraction of both property fields.

Hydropathy is more strongly aligned with the codon-position coordinates than residue volume. This agrees with the visual separation of the NUN and NAN groups along the horizontal axis.

The largest hydropathy coefficient in the complete padded analysis remains the second-position uracil term:

$$\beta_{2,U}^{(H)} \approx 3.95.$$

The corresponding historical sense-codon value is

$$\beta_{2,U}^{(H)} \approx 4.02.$$

The close agreement shows that the dominant uracil effect is robust to the inclusion of padded stop codons.

For residue volume, the second-position cytosine class is associated with smaller residues, while second-position uracil is associated with larger residues.

The combined physicochemical interpretation is therefore

$$\boxed{U \text{ at position } 2 \longrightarrow \text{hydrophobic and relatively large residues,}}$$

and

$$\boxed{C \text{ at position } 2 \longrightarrow \text{relatively small residues.}}$$

The hydropathy–volume figure and the coefficient analysis describe the same organization in complementary forms. The figure displays the grouping geometrically. The matrix analysis measures it numerically.

The central conclusion of this section is that the second codon position acts as a major organizer of amino-acid physicochemical space.

4. Complete Codon Space, Rank Ten, and the Exact Pseudoinverse

4.1 The Complete 64×12 Codon Matrix

The complete codon space contains every possible three-nucleotide word formed from the RNA alphabet

$$\mathcal{A} = \{A, G, C, U\}.$$

Because four nucleotides are available at each of three positions, the number of codons is

$$4^3 = 64.$$

Each codon can be represented by twelve binary variables: four nucleotide indicators for the first position, four for the second position, and four for the third position.

Let

$$D \in \{0,1\}^{64\times 12}$$

denote the complete codon matrix.

For codon c, position p, and nucleotide s,

$$D_{c,(p,s)} = \begin{cases} 1, & c_p = s, \\ 0, & c_p \neq s. \end{cases}$$

Each row contains exactly three entries equal to one because every codon contains one nucleotide at each of the three positions.

Using the column order A, G, C, U, the twelve columns are

$$(1, A), (1, G), (1, C), (1, U),$$
$$(2, A), (2, G), (2, C), (2, U),$$

and

$$(3, A), (3, G), (3, C), (3, U).$$

For example, the codon AUG activates the columns

$$(1, A),$$
$$(2, U),$$

and

$$(3, G).$$

Its row therefore contains three ones and nine zeros.

The matrix D describes codon composition only. It does not contain amino-acid assignments or physicochemical values. Those enter later through the codon-to-amino-acid mapping and the selected amino-acid property vectors.

This distinction is important. The matrix supplies the universal coordinate system of complete triplet space. The standard genetic code supplies the biological labeling of that space.

4.2 Complete Factorial Balance

The complete codon matrix is highly regular because all sixty-four triplets are present.

At a fixed codon position, each nucleotide occurs equally often. Once one nucleotide and one position are fixed, the other two positions can vary independently over four possibilities each.

Therefore, every nucleotide occurs

$$4^2 = 16$$

times at each codon position.

Thus, every column of D contains exactly sixteen ones.

For every position p and nucleotide s,

$$\sum_{c} D_{c,(p,s)} = 16.$$

Now consider two distinct codon positions. If one nucleotide is fixed at the first position and another nucleotide is fixed at the second position, the remaining position can take any of four values.

Therefore, every ordered nucleotide pair occurs

$$4$$

times across two specified positions.

For distinct positions $p \neq q$,

$$\sum_{c} D_{c,(p,s)} D_{c,(q,t)} = 4.$$

Within the same positional block, two different nucleotide indicators never overlap because one codon cannot contain two different nucleotides at one position.

For $s \neq t$,

$$\sum_{c} D_{c,(p,s)} D_{c,(p,t)} = 0.$$

These counting relations determine the Gram matrix

$$D^T D.$$

Let I_4 be the 4×4 identity matrix and J_4 the 4×4 all-ones matrix. Then

$$\boxed{D^T D = \begin{pmatrix} 16I_4 & 4J_4 & 4J_4 \\ 4J_4 & 16I_4 & 4J_4 \\ 4J_4 & 4J_4 & 16I_4 \end{pmatrix}.}$$

The diagonal blocks describe the four mutually exclusive nucleotide states within one codon position.

The off-diagonal blocks describe the perfectly balanced co-occurrence of nucleotide states across different positions.

The matrix also satisfies

$$D\mathbf{1}_{12} = \mathbf{31}_{64},$$

because each codon row contains three ones.

Similarly,

$$D^T\mathbf{1}_{64} = \mathbf{161}_{12},$$

because each nucleotide-position column contains sixteen ones.

These exact balance relations are the reason the complete codon matrix admits a simple analytical pseudoinverse.

4.3 Rank-Ten Structure

Although Dhas twelve columns, the columns are not independent.

Within each positional block, the four nucleotide indicators sum to the all-ones vector.

Let D_1, D_2, and D_3 denote the three 64×4 positional blocks.

Then

$$D_1\mathbf{1}_4 = \mathbf{1}_{64},$$
$$D_2\mathbf{1}_4 = \mathbf{1}_{64},$$

and

$$D_3\mathbf{1}_4 = \mathbf{1}_{64}.$$

Therefore,

$$D_1\mathbf{1}_4 - D_2\mathbf{1}_4 = 0,$$

and

$$D_1\mathbf{1}_4 - D_3\mathbf{1}_4 = 0.$$

These are two independent linear dependencies among the twelve columns.

The rank is therefore

$$\boxed{\text{rank}\,(D) = 12 - 2 = 10.}$$

The nullity is

$$\boxed{\text{nullity}\,(D) = 2.}$$

The ten independent dimensions can be interpreted as

$$\boxed{10 = 1 + 3 + 3 + 3.}$$

The first dimension is the overall mean.

Each codon position contributes three independent nucleotide contrasts because four nucleotide categories contain one shared mean direction and three contrast directions.

The additive codon property space can therefore be written as

$$\hat{y}(c) = \mu + \alpha_1(c_1) + \alpha_2(c_2) + \alpha_3(c_3).$$

Here, μ is the overall mean property, and $\alpha_p(s)$ is the centered contribution of nucleotide s at codon position p.

A convenient constraint is

$$\sum_{s \in \{A,G,C,U\}} \alpha_p\,(s) = 0.$$

Under this constraint, the additive decomposition is unique.

The rank-ten result does not depend on amino-acid assignments. Any complete four-letter triplet system represented in this way would have the same rank.

Its biological importance appears when amino-acid property fields lie strongly within this ten-dimensional subspace.

4.4 The Exact Moore–Penrose Pseudoinverse

Because D is rectangular and rank deficient, an ordinary inverse does not exist.

The appropriate generalized inverse is the Moore–Penrose pseudoinverse.

For the complete 64×12 codon matrix, the pseudoinverse has the exact analytical form

$$\boxed{D^{+} = \frac{1}{16} D^{T} - \frac{1}{96} J_{12\times 64}.}$$

This is a 12×64 matrix.

The formula contains two terms.

The first term,

$$\frac{1}{16} D^{T},$$

assigns positive weight to codons that contain a specified nucleotide at a specified position.

The second term,

$$\frac{1}{96} J,$$

subtracts a uniform correction that removes the redundancy caused by the three equal positional block sums.

Elementwise, the pseudoinverse is

$$(D^{+})_{(p,s),c} = \frac{1}{16} \mathbf{1}(c_p = s) - \frac{1}{96}.$$

Therefore, every entry has one of only two possible values.

If codon $c$contains nucleotide $s$at position p,

$$(D^{+})_{(p,s),c} = \frac{5}{96}.$$

If it does not,

$$(D^+)_{(p,s),c} = -\frac{1}{96}.$$

Thus,

$$\boxed{(D^+)_{(p,s),c} = \begin{cases} \frac{5}{96}, & c_p = s, \\ -\frac{1}{96}, & c_p \neq s. \end{cases}}$$

The entire pseudoinverse is therefore generated by one match-versus-nonmatch rule.

For any complete codon property vector

$$\mathbf{y} \in \mathbb{R}^{64},$$

the nucleotide-position coefficient vector is

$$\boxed{\boldsymbol{\beta} = D^+\mathbf{y}.}$$

The reconstructed additive property field is

$$\boxed{\hat{\mathbf{y}} = DD^+\mathbf{y}.}$$

The matrix

$$DD^+$$

is the orthogonal projection from the full sixty-four-dimensional codon property space onto the ten-dimensional additive nucleotide-position subspace.

The pseudoinverse satisfies the Moore–Penrose conditions:

$$DD^+D = D,$$
$$D^+DD^+ = D^+,$$
$$(DD^+)^T = DD^+,$$

and

$$(D^{+}D)^{T} = D^{+}D.$$

The value of the exact formula is not merely computational convenience. It reveals that the complete codon matrix has a highly regular inverse geometry determined by factorial balance.

4.5 Brief Relationship to the Historical SVD Analysis

The original hydropathy analysis used the sixty-one sense codons and excluded stop codons.

Its coefficient vector was

$$\boldsymbol{\beta}_{61} = D_{61}^{+}\mathbf{h}_{61}.$$

Because the three stop rows were removed, the exact complete-cube formula could not be applied directly. The pseudoinverse was therefore obtained numerically through singular-value decomposition.

The complete-codon formulation restores all sixty-four triplets by assigning defined values to stop X.

For hydropathy,

$$H_X = 0.$$

For residue volume,

$$V_X = \frac{1}{20}\sum_{r=1}^{20} V_r.$$

The corresponding complete-cube coefficients are

$$\boldsymbol{\beta}_{64} = D^{+}\mathbf{y}_{64}.$$

The distinction can be stated once and compactly:

the 61-codon SVD reproduces the historical biological fit,

while

the padded 64-codon formula preserves complete symmetry.

The coefficient values differ slightly because the fitted data sets are not identical.

The principal biological conclusion is unchanged. In both calculations, uracil in the second codon position is the dominant positive determinant of amino-acid hydropathy.

The remainder of the paper uses the complete codon matrix where exact symmetry is mathematically useful, without repeatedly revisiting the methodological distinction.

5. Mapping Amino-Acid Properties onto Nucleotide Positions

5.1 The Codon-to-Amino-Acid Assignment Matrix

The complete codon matrix Ddescribes the nucleotide present at each position of every codon. A separate matrix is required to describe the biological assignment of codons to amino acids and translation stop.

Let

$$\mathcal{R} = \{20 \text{ amino acids}\} \cup \{X\}$$

be the set of twenty-one translated labels, where Xdenotes stop.

Define the codon-to-label assignment matrix

$$G \in \{0,1\}^{64\times 21}.$$

For codon $c$and translated label r,

$$G_{c,r} = \begin{cases} 1, & \phi(c) = r, \\ 0, & \phi(c) \neq r. \end{cases}$$

Every codon has exactly one translated label. Therefore, every row of Gcontains one entry equal to one:

$$\sum_{r\in\mathcal{R}} G_{c,r} = 1.$$

The column sums record genetic-code degeneracy. If label $r$is assigned to $n_r$codons, then

$$\sum_c G_{c,r} = n_r.$$

For the twenty amino acids and stop,

$$\sum_{r \in \mathcal{R}} n_r = 64.$$

The matrix G is an algebraic representation of the genetic-code table. It contains the codon assignments but no physicochemical information.

The matrix D, by contrast, contains the nucleotide-position coordinates but no amino-acid assignments.

The complete mapping problem therefore requires both matrices:

$$D = \text{codon coordinate structure},$$

and

$$G = \text{codon-to-amino-acid assignment}.$$

Their roles are distinct. The matrix D is unchanged if the amino-acid assignments are randomized. The matrix G changes whenever the genetic code changes.

5.2 Amino-Acid Property Vectors and Matrices

Let a numerical property be assigned to each amino acid. Examples include hydropathy, residue volume, polarity, charge, molecular mass, aromaticity, and conformational preference.

For one property, define

$$\mathbf{q} \in \mathbb{R}^{21}.$$

The first twenty entries contain the amino-acid values, and the final entry contains the chosen padding value for stop X.

For Kyte–Doolittle hydropathy,

$$\mathbf{q} = \mathbf{h}.$$

The stop value is

$$h_X = 0.$$

For amino-acid residue volume,

$$\mathbf{q} = \mathbf{v}.$$

The stop value is the mean of the twenty amino-acid volumes:

$$v_X = \frac{1}{20}\sum_{r=1}^{20} v_r.$$

For the scale used in this analysis,

$$v_X \approx 141.26\ \text{Å}^3.$$

Several properties can be analyzed simultaneously. Let

$$Q \in \mathbb{R}^{21\times m}$$

contain m property vectors as columns.

For hydropathy and residue volume,

$$Q = (\mathbf{h} \quad \mathbf{v}).$$

Each row of Q describes one amino acid or stop X, while each column describes one physicochemical property.

This representation makes the method naturally multivariate. The same matrix operation that maps hydropathy onto nucleotide positions can map residue volume or any collection of amino-acid properties.

5.3 From Amino Acids to Codon Property Fields

The assignment matrix transfers amino-acid properties onto the codons that encode them.

For one property vector,

$$\boxed{\mathbf{y} = G\mathbf{q}.}$$

The vector

$$\mathbf{y} \in \mathbb{R}^{64}$$

contains one property value for every codon.

If codon $c$encodes amino acid r, then

$$y_c = q_r.$$

Synonymous codons therefore receive identical values.

For hydropathy,

$$\boxed{\mathbf{h}_{64} = G\mathbf{h}.}$$

For residue volume,

$$\boxed{\mathbf{v}_{64} = G\mathbf{v}.}$$

For several properties,

$$\boxed{Y = GQ.}$$

The matrix

$$Y \in \mathbb{R}^{64\times m}$$

is a collection of physicochemical fields over complete codon space.

For the paired analysis,

$$Y = (\mathbf{h}_{64} \quad \mathbf{v}_{64}).$$

This step converts a property defined on twenty amino acids into a property defined on sixty-four codons.

Degeneracy is automatically represented. An amino acid encoded by six codons contributes its value to six rows of Y, while an amino acid encoded by one codon contributes its value to one row.

The codon-level field therefore reflects both amino-acid chemistry and genetic-code degeneracy.

The transformation

$$Q \longrightarrow GQ$$

is the first stage of the analysis:

$$\boxed{\text{amino-acid property space} \longrightarrow \text{codon property space.}}$$

No fitting occurs at this stage. The property values are simply transferred according to the standard genetic-code assignments.

5.4 From Codon Fields to Twelve Nucleotide Coefficients

The next step maps the codon-level property fields onto the twelve nucleotide-position variables.

The coefficient matrix is

$$\boxed{B = D^{+}Y.}$$

Substituting

$$Y = GQ$$

gives the central mapping equation:

$$\boxed{B = D^{+}GQ.}$$

The dimensions are

$$D^{+} \in \mathbb{R}^{12\times 64},$$
$$G \in \mathbb{R}^{64\times 21},$$

and

$$Q \in \mathbb{R}^{21\times m}.$$

Therefore,

$$B \in \mathbb{R}^{12\times m}.$$

Each row of B represents one nucleotide-position state. Each column represents one amino-acid property.

For the two-property analysis,

$$B = (\boldsymbol{\beta}^{(H)} \quad \boldsymbol{\beta}^{(V)}).$$

The first column contains the twelve hydropathy coefficients. The second contains the twelve residue-volume coefficients.

For one codon property vector,

$$\boxed{\boldsymbol{\beta} = D^{+}\mathbf{y}.}$$

Using the exact pseudoinverse,

$$D^{+} = \frac{1}{16}D^{T} - \frac{1}{96}J,$$

the coefficient vector is

$$\boldsymbol{\beta} = \frac{1}{16}D^{T}\mathbf{y} - \frac{1}{96}J\mathbf{y}.$$

To interpret this result, define the set of codons containing nucleotide $s$at position p:

$$\mathcal{C}_{p,s} = \{c: c_p = s\}.$$

Every such set contains sixteen codons:

$$|\ \mathcal{C}_{p,s}\ | = 16.$$

Define the conditional mean property

$$\bar{y}_{p,s} = \frac{1}{16} \sum_{c \in \mathcal{C}_{p,s}} y_c.$$

Define the overall codon mean

$$\mu = \frac{1}{64} \sum_{c} y_c.$$

The exact nucleotide-position coefficient is then

$$\boxed{\beta_{p,s} = \bar{y}_{p,s} - \frac{2}{3}\mu.}$$

Thus, each coefficient is determined by two quantities:

the mean property of codons containing s at position p,

and

the global mean property across all codons.

The correction term appears because the predicted value of every codon contains three nucleotide coefficients, one from each positional block.

The exact formula makes the property mapping transparent. A nucleotide receives a large positive coefficient when codons containing that nucleotide at that position tend to encode amino acids with property values above the global mean.

It receives a negative coefficient when the associated codons tend to encode values below the global mean.

For hydropathy, codons containing U in the second position have a high conditional mean. Consequently,

$$\beta_{2,U}^{(H)}$$

is strongly positive.

Codons containing A in the second position have a low hydropathy mean. Consequently,

$$\beta_{2,A}^{(H)}$$

is strongly negative.

For residue volume, codons containing C in the second position tend to encode smaller amino acids. Their conditional volume mean is therefore below the global mean.

The coefficient map quantitatively reproduces the grouping visible in the hydropathy–volume figure.

5.5 Reconstruction from Nucleotide Contributions

Once the nucleotide-position coefficients have been calculated, the codon property fields are reconstructed through

$$\boxed{\hat{Y} = DB.}$$

Substituting the property map gives

$$\boxed{\hat{Y} = DD^{+}GQ.}$$

For one property,

$$\boxed{\hat{\mathbf{y}} = DD^{+}\mathbf{y}.}$$

The predicted property associated with codon

$$c = (c_1, c_2, c_3)$$

is

$$\boxed{\hat{y}(c) = \beta_{1,c_1} + \beta_{2,c_2} + \beta_{3,c_3}.}$$

The minimum-norm coefficients distribute the global mean among the three codon positions. A centered representation is often more biologically intuitive.

Define the centered nucleotide effect

$$\alpha_p(s) = \bar{y}_{p,s} - \mu.$$

The centered effects satisfy

$$\sum_s \alpha_p\,(s) = 0$$

for each codon position.

The reconstructed property can then be written as

$$\boxed{\hat{y}(c) = \mu + \alpha_1(c_1) + \alpha_2(c_2) + \alpha_3(c_3).}$$

This form separates the global property mean from the effects of the individual nucleotides.

An equivalent expression uses the three conditional means:

$$\boxed{\hat{y}(c) = \bar{y}_{1,c_1} + \bar{y}_{2,c_2} + \bar{y}_{3,c_3} - 2\mu.}$$

The prediction is therefore obtained from the average properties associated with the three nucleotide-position states present in the codon.

The residual is

$$\boxed{\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}.}$$

Equivalently,

$$\boxed{\mathbf{r} = (I - DD^{+})\mathbf{y}.}$$

The residual contains the physicochemical structure not represented by independent additive nucleotide effects.

The projected and residual components are orthogonal:

$$\hat{\mathbf{y}}^T\mathbf{r} = 0.$$

Therefore,

$$\boxed{\| \mathbf{y} \|^2 = \| \hat{\mathbf{y}} \|^2 + \| \mathbf{r} \|^2.}$$

For a centered property field, the explained variance is

$$R^2 = 1 - \frac{\| \mathbf{r} \|^2}{\| \mathbf{y} - \mu \mathbf{1} \|^2}.$$

The hydropathy field projects more strongly into the additive nucleotide-position space than the residue-volume field, but both show substantial organization.

5.6 Biological Interpretation

The mapping

$$\boxed{B = D^{+} G Q}$$

is the central bridge between amino-acid chemistry and codon composition.

The matrix Q contains the physical properties of the amino acids.

The matrix G places those properties onto the codons according to the standard genetic code.

The pseudoinverse D^{+}expresses the resulting codon property field in terms of nucleotide identity and codon position.

The complete transformation is

$$\boxed{Q \overset{G}{\rightarrow} Y \overset{D^{+}}{\rightarrow} B \overset{D}{\rightarrow} \hat{Y}.}$$

The coefficients should be interpreted as statistical contributions within an additive codon model. They are not binding energies, molecular forces, or intrinsic physical properties of isolated nucleotides.

A coefficient answers a specific question:

$$\boxed{\text{What property tendency is associated with nucleotide } s \text{ at codon position } p?}$$

For hydropathy, the dominant answer is

$$\boxed{U \text{ at position } 2 \longrightarrow \text{strong hydrophobic tendency.}}$$

The complementary result is

$$\boxed{A \text{ at position } 2 \longrightarrow \text{strong hydrophilic tendency.}}$$

For residue volume,

$$\boxed{C \text{ at position } 2 \longrightarrow \text{small-residue tendency.}}$$

These effects are not equally distributed across the three codon positions. The middle nucleotide carries the strongest physicochemical classification, while the third position contributes less because many third-position substitutions are synonymous.

The mapping also reveals an important distinction between amino-acid identity and amino-acid property. The model does not usually reconstruct the exact amino-acid label from the codon coefficients. Instead, it reconstructs broad physicochemical tendencies.

This is why codons encoding different amino acids can occupy similar regions of the projected property space.

The method is extensible. Any numerical amino-acid property can be inserted as a column of Q, producing a corresponding column of nucleotide-position coefficients in B.

The same framework can therefore compare:

hydropathy,
residue volume,
charge,
polarity,

and

other physicochemical descriptors.

The central biological finding is that the standard genetic code places major amino-acid properties in a form that is substantially recoverable from nucleotide-position coordinates.

The next section shows that the projection performing this reconstruction has an exact geometric interpretation in terms of codon Hamming distance.

6. Hamming Geometry of the Codon Projection

6.1 The Projection Operator

The complete codon matrix maps twelve nucleotide-position coefficients onto a property field over the sixty-four codons:

$$\hat{\mathbf{y}} = D\boldsymbol{\beta}.$$

Using the Moore–Penrose pseudoinverse,

$$\boldsymbol{\beta} = D^{+}\mathbf{y}.$$

Substitution gives

$$\boxed{\hat{\mathbf{y}} = DD^{+}\mathbf{y}.}$$

Define the 64×64 projection matrix

$$\boxed{H = DD^{+}.}$$

The matrix H maps any codon-level property vector onto the ten-dimensional subspace that can be represented by independent nucleotide effects at positions one, two, and three.

Because H is an orthogonal projection,

$$\boxed{H^2 = H.}$$

Projecting a property field twice therefore gives the same result as projecting it once.

The projection is also symmetric:

$$\boxed{H^T = H.}$$

Its rank equals the rank of the codon matrix:

$$\boxed{\text{rank}\,(H) = 10.}$$

Thus, Hextracts the component of a sixty-four-dimensional codon property field that lies in the rank-ten additive nucleotide-position space.

The complementary projection is

$$I_{64} - H.$$

For any codon property vector,

$$\mathbf{y} = H\mathbf{y} + (I_{64} - H)\mathbf{y}.$$

The first term is the nucleotide-recoverable component. The second is the residual component requiring structure beyond independent positional effects.

6.2 Matching Positions and Hamming Distance

Let

$$c = (c_1, c_2, c_3)$$

and

$$d = (d_1, d_2, d_3)$$

be two codons.

Define the number of matching positions as

$$\boxed{m(c,d) = \sum_{p=1}^{3} \mathbf{1}(\, c_p = d_p).}$$

The possible values are

$$m(c,d) \in \{0,1,2,3\}.$$

Identical codons have

$$m(c,d) = 3.$$

Codons differing at one position have

$$m(c,d) = 2.$$

Codons differing at two positions have

$$m(c,d) = 1.$$

Codons differing at all three positions have

$$m(c,d) = 0.$$

The Hamming distance is the number of positions at which two codons differ:

$$\boxed{d_H(c,d) = \sum_{p=1}^{3} \mathbf{1}(c_p \neq d_p).}$$

Therefore,

$$\boxed{d_H(c,d) = 3 - m(c,d).}$$

Hamming distance is the natural order-independent distance on complete codon space. It does not require the nucleotides to be arranged in a particular sequence.

Two codons with

$$d_H(c,d) = 1$$

are related by one nucleotide substitution.

Each codon has three positions at which a substitution can occur. At each position, the existing nucleotide can be replaced by any of the other three nucleotides. Every codon therefore has

$$3(4-1) = 9$$

Hamming-distance-one neighbors.

The full codon Hamming graph consequently contains

$$\boxed{\frac{64 \times 9}{2} = 288}$$

undirected single-substitution edges.

This graph treats all three possible substitutions at a given position as equivalent nearest-neighbor changes.

6.3 The Exact Hamming-Distance Kernel

The entries of the projection matrix H have a simple exact relationship to Hamming distance.

The complete-codon pseudoinverse is

$$D^{+} = \frac{1}{16} D^{T} - \frac{1}{96} J.$$

Therefore,

$$H = DD^{+}.$$

For codons $c$and d, the corresponding projection coefficient is

$$H_{cd} = \frac{1}{16} (DD^{T})_{cd} - \frac{1}{96} (DJ)_{cd}.$$

The inner product of codon rows c and d equals the number of positions at which the codons contain the same nucleotide:

$$(DD^{T})_{cd} = m(c, d).$$

Every row of D contains three ones, so

$$(DJ)_{cd} = 3.$$

The projection coefficient is therefore

$$\boxed{H_{cd} = \frac{m(c,d)}{16} - \frac{1}{32}.}$$

Using

$$m(c,d) = 3 - d_H(c,d),$$

the same result becomes

$$\boxed{H_{cd} = \frac{5}{32} - \frac{d_H(c,d)}{16}.}$$

Thus, the projection coefficient between two codons depends only on their Hamming distance.

The four possible values are

$$\boxed{H_{cd} = \begin{cases} \frac{5}{32}, & d_H(c,d) = 0, \\ \frac{3}{32}, & d_H(c,d) = 1, \\ \frac{1}{32}, & d_H(c,d) = 2, \\ -\frac{1}{32}, & d_H(c,d) = 3. \end{cases}}$$

Identical codons receive the largest positive coefficient.

Single-substitution neighbors receive

$$\frac{3}{32}.$$

Codons differing at two positions receive

$$\frac{1}{32}.$$

Codons differing at all three positions receive

$$-\frac{1}{32}.$$

The diagonal entries are all equal:

$$H_{cc} = \frac{5}{32}.$$

The trace is therefore

$$\mathrm{tr}\,(H) = 64\left(\frac{5}{32}\right).$$

Hence,

$$\boxed{\mathrm{tr}\,(H) = 10.}$$

For an orthogonal projection, the trace equals the rank, confirming

$$\mathrm{rank}\,(H) = 10.$$

The row sums also equal one. For any fixed codon, there is one codon at distance zero, nine at distance one, twenty-seven at distance two, and twenty-seven at distance three.

Thus,

$$\frac{5}{32} + 9\left(\frac{3}{32}\right) + 27\left(\frac{1}{32}\right) - 27\left(\frac{1}{32}\right) = 1.$$

This confirms that the constant property field is preserved by the projection.

The predicted value of codon $c$may therefore be written as a weighted sum over all codons:

$$\boxed{\hat{y}(c) = \sum_{d} H_{cd}\, y(d).}$$

The weights depend only on the symbolic distance between c and d.

This result connects the exact pseudoinverse directly to the geometry of codon space.

6.4 Recoverable and Residual Property Structure

For a codon-level property vector $\mathbf{y}$, define the recoverable component

$$\boxed{\hat{\mathbf{y}} = H\mathbf{y}.}$$

Define the residual component

$$\boxed{\mathbf{r} = (I_{64} - H)\mathbf{y}.}$$

The two components are orthogonal:

$$\boxed{\hat{\mathbf{y}}^T\mathbf{r} = 0.}$$

Therefore,

$$\boxed{\| \mathbf{y} \|^2 = \| \hat{\mathbf{y}} \|^2 + \| \mathbf{r} \|^2.}$$

The projected field contains the part of hydropathy, residue volume, or another amino-acid property that can be represented by independent effects of the three codon positions.

The residual contains the property structure that cannot be written in the additive form

$$\mu + \alpha_1(c_1) + \alpha_2(c_2) + \alpha_3(c_3).$$

This residual includes interactions between positions.

For example, a special property associated only with a particular combination of first- and second-position nucleotides cannot generally be represented by separate first- and second-position contributions.

The complete sixty-four-dimensional codon function space can be decomposed into:

1

overall mean dimension,

9

single-position dimensions,

$$27$$

two-position interaction dimensions,

and

$$27$$

three-position interaction dimensions.

Thus,

$$\boxed{64 = 1 + 9 + 27 + 27.}$$

The additive nucleotide-position model contains only the mean and the nine single-position effects:

$$\boxed{10 = 1 + 9.}$$

The remaining

$$54$$

dimensions contain interaction structure.

The significance of the physicochemical results lies in the amount of hydropathy and residue-volume variation captured by this restricted ten-dimensional space.

Hydropathy projects particularly strongly because the second codon position partitions much of the hydrophobic and hydrophilic variation.

Residue volume also has a substantial recoverable component, especially through the association of second-position cytosine with small residues.

The residual is not necessarily error or biological noise. It may contain meaningful codon interactions, irregularities of degeneracy, and distinctions among amino acids that cannot be reduced to independent positional contributions.

6.5 Hamming Graph Versus Ordered Cube

The Hamming projection is order independent. It depends only on whether nucleotide symbols match at each position.

The later cube analysis introduces an explicit linear order among the four nucleotides.

This distinction produces two different codon graphs.

The full Hamming graph connects every pair of codons differing at exactly one position. It contains

$$\boxed{288}$$

undirected edges.

The ordered $4 \times 4 \times 4$ cube connects only nucleotide states that are consecutive along the selected axis path.

For a nucleotide order

$$\pi = (\pi_1, \pi_2, \pi_3, \pi_4),$$

the adjacent nucleotide pairs are

$$\pi_1 \leftrightarrow \pi_2,$$
$$\pi_2 \leftrightarrow \pi_3,$$

and

$$\pi_3 \leftrightarrow \pi_4.$$

Each axis contributes forty-eight nearest-neighbor edges, giving

$$\boxed{3 \times 48 = 144}$$

ordered-cube edges.

The relationship can therefore be summarized as

$$\boxed{\text{Hamming graph} = \text{all one-position substitutions},}$$

while

$$\boxed{\text{ordered cube} = \text{adjacent substitutions under } \pi.}$$

The Hamming projection kernel is invariant under nucleotide relabeling. If the four nucleotide symbols are permuted, the Hamming distances remain unchanged.

The ordered cube is not invariant under such relabeling because a different permutation changes which nucleotide pairs are adjacent.

The two geometries therefore answer different questions.

The Hamming analysis asks:

$$\boxed{\text{How is an additive codon property determined by symbolic distance?}}$$

The ordered-cube analysis asks:

$$\boxed{\text{Which nucleotide path produces the greatest synonymous adjacency?}}$$

The first describes the universal geometry of complete triplet space.

The second evaluates the topology created by the standard amino-acid assignments under alternative nucleotide orders.

This distinction prepares the transition from codon projection to the isotropic G-A-C-U cube.

7. The Isotropic G-A-C-U Codon Cube

7.1 The $4 \times 4 \times 4$ Codon Lattice

The complete codon space contains every ordered triplet formed from the four RNA nucleotides:

$$\mathcal{A} = \{A, G, C, U\}.$$

The codon set is the Cartesian product

$$\mathcal{C} = \mathcal{A}^3.$$

Therefore,

$$| \mathcal{C} | = 4^3 = 64.$$

Each codon

$$c = (c_1, c_2, c_3)$$

can be represented as one point in a three-dimensional lattice. The first coordinate represents the first codon position, the second coordinate represents the second codon position, and the third coordinate represents the third codon position.

Because each coordinate can assume four nucleotide states, the resulting geometry is a

$$4 \times 4 \times 4$$

lattice.

To place the nucleotides along a numerical axis, a linear nucleotide order must be selected. Let

$$\pi = (\pi_0, \pi_1, \pi_2, \pi_3)$$

be a permutation of the four nucleotides.

Define the coordinate function

$$\kappa_\pi(\pi_i) = i.$$

Thus, the first nucleotide in the selected order receives coordinate 0, the second receives coordinate 1, the third receives coordinate 2, and the fourth receives coordinate 3.

The codon coordinate map is

$$\boxed{\mathbf{x}_\pi(c) = (\kappa_\pi(c_1), \kappa_\pi(c_2), \kappa_\pi(c_3)).}$$

Every codon is assigned a unique lattice point in

$$\{0, 1, 2, 3\}^3.$$

Conversely, every lattice point corresponds to one codon. The mapping is therefore one-to-one.

The cube preserves the ordered nature of codons. Exchanging the first and second coordinates changes the codon because codon positions are biologically distinct.

The three lattice axes have direct interpretations:

$$\text{axis 1} \longleftrightarrow \text{first codon position},$$
$$\text{axis 2} \longleftrightarrow \text{second codon position},$$

and

$$\text{axis 3} \longleftrightarrow \text{third codon position}.$$

A movement parallel to axis one changes only the first nucleotide. A movement parallel to axis two changes only the second nucleotide. A movement parallel to axis three changes only the third nucleotide.

This geometric representation makes positional structure visible. In particular, planes perpendicular to axis two contain codons sharing the same middle nucleotide. These planes correspond to the four physicochemical families discussed previously.

7.2 Isotropic Nucleotide Ordering

A general codon lattice could use a different nucleotide order on each axis. Let

$$\pi_1, \pi_2, \pi_3$$

denote the orders assigned to codon positions one, two, and three.

Such a lattice is anisotropic when

$$\pi_1, \pi_2, \pi_3$$

are allowed to differ.

The present analysis first considers isotropic cubes. An isotropic cube uses one common nucleotide order on all three axes:

$$\boxed{\pi_1 = \pi_2 = \pi_3 = \pi.}$$

The same nucleotide path therefore governs substitutions at all three codon positions.

This condition provides a controlled and symmetric comparison among nucleotide orders. Every candidate order is applied identically to the first, second, and third codon axes.

The isotropic cube does not assume that the three codon positions are biologically equivalent. Their amino-acid effects remain different because the standard genetic-code assignments are position dependent.

Isotropy refers only to the geometrical ordering of the nucleotide states.

For a chosen order

$$\pi = (\pi_0, \pi_1, \pi_2, \pi_3),$$

the consecutive nucleotide pairs are

$$\pi_0 \leftrightarrow \pi_1,$$
$$\pi_1 \leftrightarrow \pi_2,$$

and

$$\pi_2 \leftrightarrow \pi_3.$$

Only these consecutive pairs are connected directly along a cube axis.

For example, if

$$\pi = (G, A, C, U),$$

then the adjacent nucleotide pairs are

$$G \leftrightarrow A,$$
$$A \leftrightarrow C,$$

and

$$C \leftrightarrow U.$$

The pairs

$$G \leftrightarrow C,$$
$$G \leftrightarrow U,$$

and

$$A \leftrightarrow U$$

are not direct nearest neighbors in this ordered cube.

This differs from the full Hamming graph, in which every nucleotide can be replaced directly by any of the other three nucleotides at one codon position.

The ordered cube therefore selects a structured subset of the full single-substitution graph.

7.3 The Twenty-Four Isotropic Permutations

The four nucleotides can be placed in a linear order in

$$4!$$

different ways.

Therefore, the number of isotropic nucleotide orders is

$$\boxed{|\ S_4\ | = 4! = 24.}$$

Each permutation

$$\pi \in S_4$$

defines one ordered codon cube.

The codon set remains fixed:

$$\mathcal{C} = \{A, G, C, U\}^3.$$

The amino-acid assignment also remains fixed:

$$\phi: \mathcal{C} \to \mathcal{R}.$$

Only the nearest-neighbor edge set changes.

For nucleotide order π, define the codon graph

$$G_\pi = (\mathcal{C}, E_\pi).$$

The vertex set contains the same sixty-four codons for every permutation. The edge set E_π contains the codon pairs that are nearest neighbors under the selected nucleotide path.

Thus, the exhaustive comparison holds the biological genetic code constant while varying the geometrical embedding.

The comparison asks:

Which common nucleotide order places the greatest number of synonymous codons next to one another?

The twenty-four permutations include reversal pairs. If

$$\pi = (\pi_0, \pi_1, \pi_2, \pi_3),$$

define its reversal as

$$R\pi = (\pi_3, \pi_2, \pi_1, \pi_0).$$

Reversal preserves the same undirected consecutive nucleotide pairs. It changes their orientation but not their adjacency.

For example,

$$R(G, A, C, U) = (U, C, A, G).$$

The G-A-C-U path contains the undirected pairs

$$\{G, A\}, \{A, C\}, \{C, U\}.$$

The reversed U-C-A-G path contains

$$\{U, C\}, \{C, A\}, \{A, G\}.$$

These are the same three undirected pairs.

Consequently,

$$\boxed{E_\pi = E_{R\pi}}$$

for the undirected cube graph.

The twenty-four oriented orders therefore correspond to twelve distinct undirected adjacency patterns.

The exhaustive analysis evaluates all twenty-four orders, while the final interpretation recognizes reversal-equivalent pairs as mirror forms of one geometric solution.

7.4 The G-A-C-U Coordinate System

The nucleotide order of principal interest is

$$\boxed{\pi_{\mathrm{GACU}} = (G, A, C, U).}$$

Its coordinate function is

$$\kappa(G) = 0,$$
$$\kappa(A) = 1,$$
$$\kappa(C) = 2,$$

and

$$\kappa(U) = 3.$$

The coordinate of codon

$$c = (c_1, c_2, c_3)$$

is therefore

$$\boxed{\mathbf{x}_{\mathrm{GACU}}(c) = (\kappa(c_1), \kappa(c_2), \kappa(c_3)).}$$

For example, the codon

$$GAC$$

has coordinate

$$\mathbf{x}_{\mathrm{GACU}}(GAC) = (0,1,2).$$

The codon

$$AUG$$

has coordinate

$$\mathbf{x}_{\mathrm{GACU}}(AUG) = (1,3,0).$$

The three nucleotide transitions represented by nearest-neighbor movements are

$$G \leftrightarrow A,$$
$$A \leftrightarrow C,$$

and

$$C \leftrightarrow U.$$

This same path is used on each codon-position axis.

The second-axis planes appear in the sequence

$$NGN,$$
$$NAN,$$
$$NCN,$$

and

$$NUN.$$

Thus, the G-A-C-U coordinate system arranges the four second-position physicochemical families in the order

$$G \to A \to C \to U.$$

The NUN plane lies at one end of the second-position axis and contains the strongly hydrophobic codon family.

The NAN plane contains many hydrophilic residues.

The NCN plane contains relatively small residues.

The NGN plane is more heterogeneous.

The cube therefore provides a geometric setting in which both physicochemical organization and synonymous connectivity can be examined.

The mirror order is

$$\boxed{\pi_{\mathrm{UCAG}} = (U, C, A, G).}$$

Because it contains the same undirected nucleotide adjacencies, it generates the same codon edge set viewed through a reflected coordinate system.

The two orders should therefore be interpreted as one undirected topology with two orientations.

7.5 Figure 2: The G-A-C-U Codon Cube

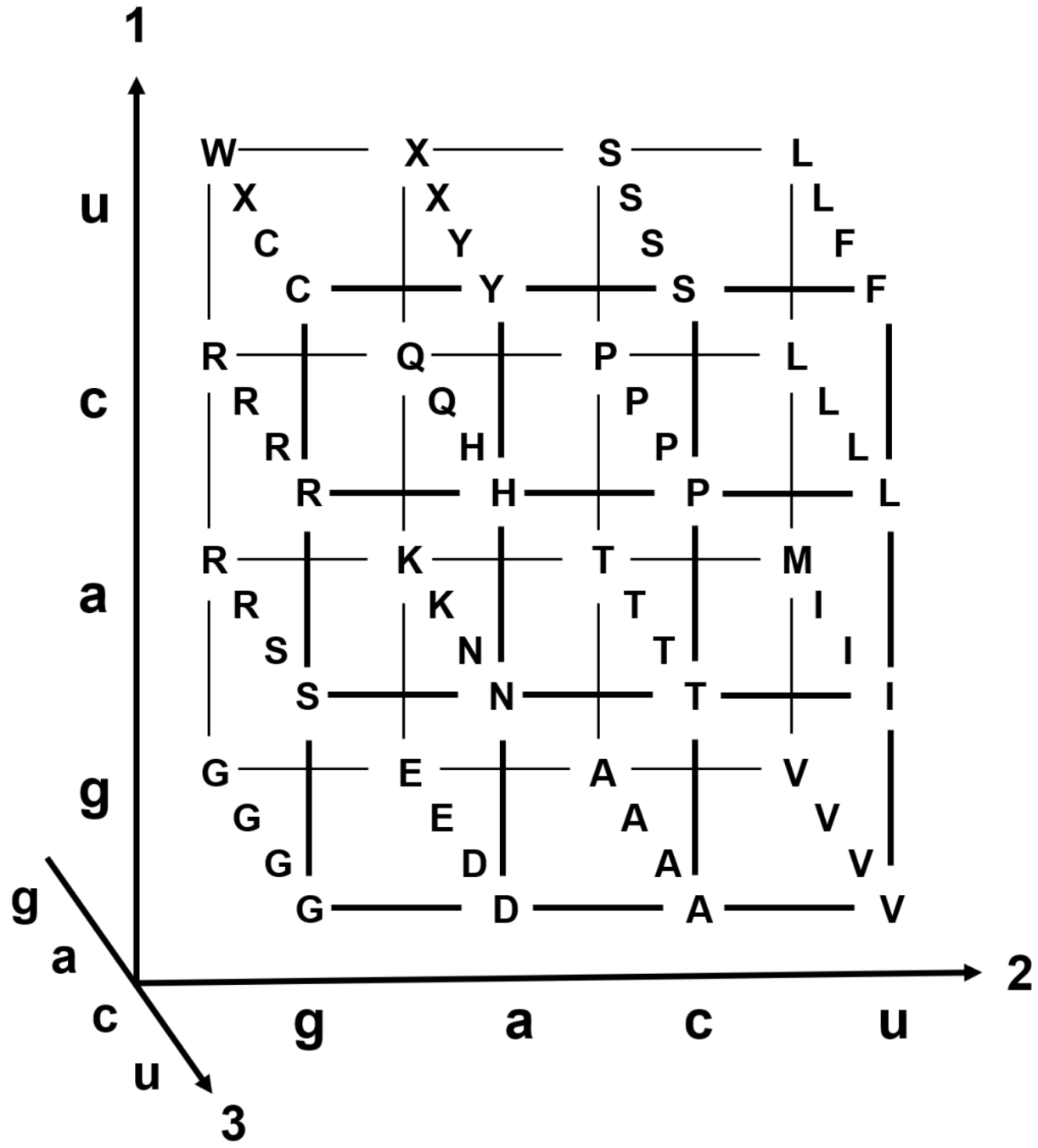

Figure 2: The G-A-C-U Codon Cube The isotropic G-A-C-U codon cube. Each lattice site represents one codon. The visible letter at that site is the single-letter code of the encoded amino acid, with X representing translation stop.

The axes correspond to codon positions one, two, and three. Each axis is ordered

$$G - A - C - U.$$

Lines parallel to axis three often connect synonymous codons because third-position variation frequently preserves amino-acid identity.

Lines parallel to axis two move between the physicochemical classes defined by the second nucleotide.

Lines parallel to axis one often refine amino-acid identity within broader second-position classes.

The cube therefore displays several biological structures simultaneously:

codon position,
synonymous degeneracy,
amino-acid identity,

and

physicochemical grouping.

The lattice itself contains sixty-four codon vertices. The next stage of the analysis collapses these vertices to twenty amino acids and stop X, while retaining the cube edges. That collapse produces the amino-acid multigraph in which synonymous connections become self-loops.

7.6 Relationship Between the Cube and the Codon Matrix

The 64×12codon matrix and the $4 \times 4 \times 4$ cube represent the same complete triplet space, but they emphasize different mathematical properties.

The matrix representation is

$$D_{c,(p,s)} = \mathbf{1}(c_p = s).$$

It records whether nucleotide $s$occurs at position $p$in codon c.

The cube representation is

$$\mathbf{x}_\pi(c) = (\kappa_\pi(c_1), \kappa_\pi(c_2), \kappa_\pi(c_3)).$$

It assigns numerical coordinates according to the selected nucleotide order.

The matrix representation does not require a linear ordering of the nucleotides. Permuting the nucleotide labels simply permutes columns within the positional blocks.

Its rank remains

$$10,$$

and its projection kernel remains determined by Hamming distance.

The cube representation does require an order because its nearest-neighbor edges depend on which nucleotide states are consecutive.

This distinction can be summarized as

$$\boxed{D = \text{order-independent codon algebra,}}$$

whereas

$$\boxed{G_\pi = \text{order-dependent codon topology.}}$$

The matrix asks how codon-level properties can be represented through nucleotide-position contributions.

The cube asks how the translated labels are arranged under a selected nucleotide path.

The Hamming graph connects all codons differing at one position. It contains

$$288$$

undirected edges.

The ordered cube retains only the substitutions between consecutive nucleotide states. It contains

$$144$$

undirected edges.

Thus,

$$\boxed{E_\pi \subset E_{\text{Hamming}}.}$$

The ordered cube selects exactly half of the full one-substitution edges.

The standard amino-acid assignment labels each cube vertex through

$$\phi(c).$$

The resulting labeled graph contains the information needed for the topology analysis.

When two adjacent codons have the same translated label, their edge is synonymous.

When the codon vertices are collapsed by translated identity, that synonymous edge becomes a self-loop.

The G-A-C-U cube is important because its labeled edge structure yields the greatest number of such self-loops among all isotropic nucleotide orders.

The following section constructs the 21-vertex amino-acid multigraph and defines the self-loop count used to measure synonymous compactness.

8. Collapse to the 21-Vertex Amino-Acid Multigraph

8.1 The 144 Nearest-Neighbor Lattice Edges

The isotropic codon cube contains sixty-four codon vertices arranged on a $4 \times 4 \times 4$ lattice. Each axis represents one codon position, and each axis contains the same ordered nucleotide path.

For an isotropic order

$$\pi = (\pi_0, \pi_1, \pi_2, \pi_3),$$

the adjacent nucleotide pairs are

$$\pi_0 \leftrightarrow \pi_1,$$
$$\pi_1 \leftrightarrow \pi_2,$$

and

$$\pi_2 \leftrightarrow \pi_3.$$

Along one axis, there are therefore

$$4 - 1 = 3$$

adjacent intervals.

For each interval, the other two codon coordinates can independently take four values. The number of edges parallel to one axis is therefore

$$(4 - 1)4^2.$$

Thus,

$$(4 - 1)4^2 = 48.$$

Because there are three codon-position axes, the total number of undirected nearest-neighbor edges is

$$\boxed{| E_\pi |= 3(4 - 1)4^2.}$$

Therefore,

$$\boxed{| E_\pi |= 144.}$$

The edge count is the same for every isotropic nucleotide ordering. The order changes which nucleotide substitutions are represented by the edges, but it does not change the number of lattice edges.

The ordered codon graph is

$$G_\pi = (\mathcal{C}, E_\pi),$$

where

$$| \mathcal{C} |= 64$$

and

$$| E_\pi | = 144.$$

Two codons are connected by an edge when they differ at exactly one codon position and the two nucleotides at that position are consecutive under the selected nucleotide order.

For the G-A-C-U path, the permitted nucleotide adjacencies are

$$G \leftrightarrow A,$$
$$A \leftrightarrow C,$$

and

$$C \leftrightarrow U.$$

A codon pair differing by $G \leftrightarrow C$, $G \leftrightarrow U$, or $A \leftrightarrow U$ is not joined directly in this ordered cube.

The 144-edge graph is therefore a selected subgraph of the full codon Hamming graph.

The full Hamming graph contains all one-position nucleotide substitutions and has

$$288$$

undirected edges.

The ordered cube retains exactly half of those edges:

$$\boxed{| E_\pi | = \frac{1}{2} | E_{\text{Hamming}} |.}$$

This ordered subset makes the topology depend on the nucleotide path π.

8.2 Collapse by Translated Identity

The sixty-four codons are not translated into sixty-four different outputs. They map to twenty amino acids and the stop signal X.

Let

$$\phi: \mathcal{C} \to \mathcal{R}$$

be the standard translation map, where

$$\mathcal{R} = \{20 \text{ amino acids}\} \cup \{X\}.$$

Therefore,

$$| \mathcal{R} | = 21.$$

The codon graph can be collapsed by identifying all vertices that have the same translated label.

If codons c and d satisfy

$$\phi(c) = \phi(d),$$

then they are mapped to the same vertex in the collapsed graph.

The resulting graph has twenty-one vertices:

$$\boxed{| V_{\text{collapsed}} | = 21.}$$

These vertices are labeled by the twenty amino-acid single-letter codes and X.

The collapse can be viewed as a quotient operation on the codon graph. Codons are partitioned into equivalence classes defined by

$$c \sim d$$

whenever

$$\phi(c) = \phi(d).$$

Each equivalence class becomes one vertex.

The collapsed graph is written as

$$M_\pi = (\mathcal{R}, \mathcal{E}_\pi).$$

The vertex set changes from sixty-four codons to twenty-one translated labels, but every original codon edge is retained.

Therefore,

$$\boxed{|\ \mathcal{E}_{\pi}\ | = 144.}$$

The collapse changes the endpoints of the edges but not their total number.

The transformation is

$$\boxed{64 \text{ codon vertices} \longrightarrow 21 \text{ translated vertices.}}$$

At the same time,

$$\boxed{144 \text{ codon edges} \longrightarrow 144 \text{ multigraph edges.}}$$

Each codon edge

$$e = \{c, d\}$$

is mapped to the translated edge

$$\phi_*(e) = \{\phi(c), \phi(d)\}.$$

If the two codons encode different amino acids, the translated edge joins two different vertices.

If the two codons encode the same amino acid, the translated edge begins and ends at the same vertex.

This second case produces a self-loop.

8.3 Parallel Edges

The collapsed graph is not a simple graph. It is a multigraph.

In a simple graph, only one edge is allowed between two vertices. In the amino-acid multigraph, several different codon pairs can connect the same pair of translated labels.

Suppose codon pair

$$\{c_1, d_1\}$$

and codon pair

$$\{c_2, d_2\}$$

are both nearest neighbors in the ordered cube.

If

$$\phi(c_1) = \phi(c_2) = r$$

and

$$\phi(d_1) = \phi(d_2) = s,$$

then both codon edges become edges between translated vertices $r$and s.

The multigraph therefore retains multiplicity.

Define

$$m_\pi(r, s)$$

as the number of codon-level edges that collapse to the translated pair (r, s).

For distinct translated labels,

$$r \neq s,$$

the value

$$m_\pi(r, s)$$

is the number of parallel edges connecting r and s.

For identical translated labels,

$$r = s,$$

the value

$$m_\pi(r, r)$$

is the number of self-loops at vertex r.

The multiplicities satisfy

$$\boxed{\sum_{r \leq s} m_\pi(r,s) = 144.}$$

The exact notation used for the unordered sum is less important than the principle: every codon edge contributes once to one translated edge class.

Parallel edges preserve biological information that would be lost if the graph were simplified.

A thick or repeated connection between two amino-acid vertices means that several distinct codon substitutions connect those translated outputs.

A single edge means that only one codon pair contributes to that translated relationship.

The multigraph therefore represents not only which amino-acid transitions are possible within the ordered cube, but also how many codon-level routes support each transition.

The visual density of the graph reflects this multiplicity.

The vertex positions in the two-dimensional drawing are produced only for readability. They do not determine the graph topology.

The topological information lies in:

which vertices are connected,
how many parallel edges connect them,

and

how many self-loops occur at each vertex.

8.4 Synonymous Adjacencies as Self-Loops

A self-loop is created when two neighboring codons translate to the same label.

Let

$$e = \{c, d\}$$

be one nearest-neighbor codon edge.

The edge is synonymous when

$$\phi(c) = \phi(d).$$

After collapse, both endpoints become the same translated vertex.

Thus,

$$\{\phi(c), \phi(d)\} = \{r, r\}.$$

This is a self-loop at vertex r.

A synonymous self-loop therefore represents a local movement through the ordered codon cube that preserves the translated output.

Biologically, the edge corresponds to a nucleotide substitution between adjacent states of the selected axis order that does not change the encoded amino acid or stop signal.

The number of self-loops at translated vertex r is

$$\ell_r(\pi) = m_\pi(r, r).$$

The total number of self-loops is the sum over all twenty-one translated labels.

The loops are not artificial additions to the graph. They are ordinary codon edges whose endpoints become identical after quotienting by translated identity.

The number of loops contributed by an amino acid depends on both its degeneracy and the geometric arrangement of its codons.

An amino acid encoded by many codons has greater potential to produce loops, but high degeneracy alone does not determine the loop count.

Its codons must also occupy adjacent positions under the selected nucleotide order.

For example, a four-codon synonymous family arranged consecutively along one cube line can produce three adjacent synonymous edges:

$$c_1 - c_2 - c_3 - c_4.$$

After collapse, these become three self-loops at one amino-acid vertex.

Thus,

$$\boxed{\text{degeneracy} \neq \text{topological compactness.}}$$

Degeneracy counts how many codons encode an amino acid.

Topological compactness measures how those codons are connected in the selected cube.

The multigraph makes this distinction visible.

8.5 A Word-Friendly Loop-Count Definition

To avoid an overly complicated summation expression, first define a synonymous-edge indicator.

For codon edge

$$e = \{c, d\},$$

define

$$\boxed{I_\phi(e) = \mathbf{1}[\phi(c) = \phi(d)].}$$

Equivalently,

$$I_\phi(e) = \begin{cases} 1, & \phi(c) = \phi(d), \\ 0, & \phi(c) \neq \phi(d). \end{cases}$$

The total loop count for nucleotide order π is then

$$\boxed{L(\pi) = \sum_{e \in E_\pi} I_\phi\,(e).}$$

This is the principal graph statistic used in the isotropic permutation analysis.

The same quantity can be written as a sum over translated vertices:

$$\boxed{L(\pi) = \sum_{r \in \mathcal{R}} \ell_r\,(\pi).}$$

The first form counts synonymous codon edges directly.

The second form sums the loops visible at the twenty-one multigraph vertices.

The loop fraction is

$$\boxed{f_{\mathrm{loop}}(\pi) = \frac{L(\pi)}{144}.}$$

For the G-A-C-U order,

$$L(GACU) = 44.$$

Therefore,

$$f_{\mathrm{loop}}(GACU) = \frac{44}{144}.$$

Simplifying,

$$\boxed{f_{\mathrm{loop}}(GACU) = \frac{11}{36}.}$$

Numerically,

$$f_{\mathrm{loop}}(GACU) \approx 0.3056.$$

Thus, slightly more than thirty percent of the ordered-cube edges connect codons with identical translated labels.

The remaining number of non-loop edges is

$$144 - 44 = 100.$$

Therefore, the G-A-C-U multigraph contains

$$\boxed{44 \text{ self-loop edges}}$$

and

$$\boxed{100 \text{ non-loop edges.}}$$

All edges are counted with multiplicity.

8.6 Figure 3: The G-A-C-U Amino-Acid Multigraph

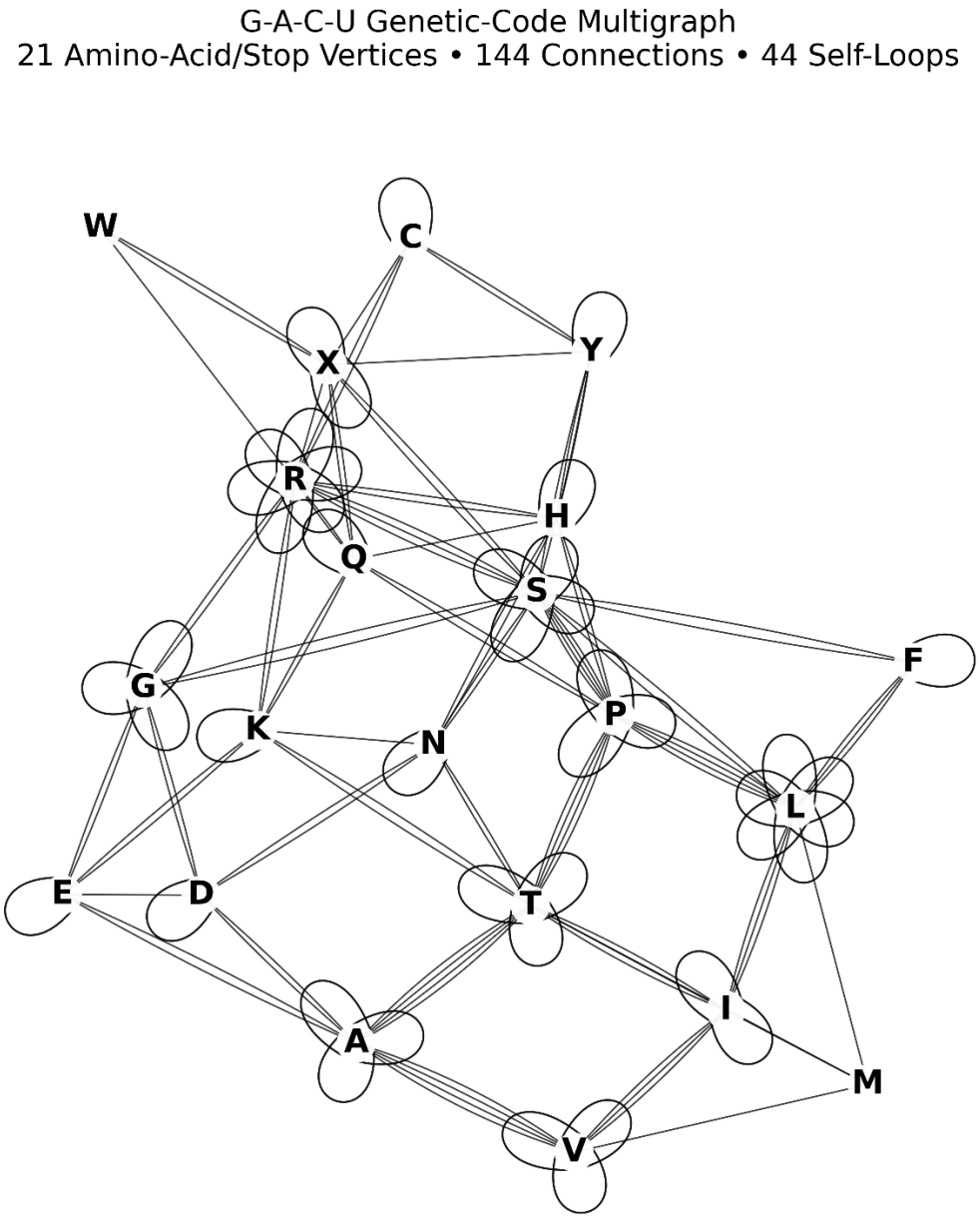

Figure 3. G-A-C-U Amino-Acid Multigraph Each labeled vertex represents one amino acid or stop X.

The letters are the standard single-letter amino-acid codes.

The curves between distinct letters represent nonsynonymous codon adjacencies.

Multiple curves between the same two letters represent parallel edges arising from different codon pairs.

The loops surrounding a letter represent synonymous nearest-neighbor codon pairs.

The graph contains

$$21$$

vertices,

$$144$$

edges in total, and

$$44$$

self-loops.

The two-dimensional placement of the vertices is only a visualization layout. Moving a letter to another location on the page would not change the graph.

The mathematically meaningful features are the edge multiplicities and self-loop counts.

The figure provides a direct visual translation of codon degeneracy into graph topology.

It also prepares the central comparison of the next section. The G-A-C-U graph is one of twenty-four isotropic graphs generated from the same codon assignments.

The exhaustive question is whether any other isotropic nucleotide order produces more than forty-four synonymous self-loops.

The answer is no.

9. Exhaustive Enumeration of Isotropic Nucleotide Orders

9.1 Computational Procedure

The isotropic search examines every possible linear ordering of the four RNA nucleotides when the same order is applied to all three codon-position axes.

The nucleotide alphabet is

$$\mathcal{A} = \{A, G, C, U\}.$$

The number of possible linear orders is

$$\boxed{4! = 24.}$$

For each permutation

$$\pi \in S_4,$$

the analysis constructs a $4 \times 4 \times 4$ codon lattice. The first, second, and third codon positions use the same ordered nucleotide path.

If

$$\pi = (\pi_0, \pi_1, \pi_2, \pi_3),$$

the adjacent nucleotide pairs are

$$\pi_0 \leftrightarrow \pi_1,$$
$$\pi_1 \leftrightarrow \pi_2,$$

and

$$\pi_2 \leftrightarrow \pi_3.$$

Every codon is placed at one lattice coordinate, and nearest-neighbor edges are drawn between codons that differ at one position by one step along the selected nucleotide path.

Each isotropic cube contains

$$64$$

codon vertices and

$$144$$

nearest-neighbor edges.

The genetic-code assignment remains fixed throughout the search. Only the nucleotide ordering changes.

The procedure for each permutation is:

1. assign numerical coordinates to A, G, C, and U;
2. place all sixty-four codons on the lattice;
3. generate the 144 nearest-neighbor edges;
4. translate the two codons at the ends of every edge;
5. count the edges whose translated labels are identical.

Using the synonymous-edge indicator defined previously,

$$I_\phi(e) = \mathbf{1}\,[\phi(c) = \phi(d)],$$

the loop count is

$$\boxed{L(\pi) = \sum_{e \in E_\pi} I_\phi\,(e).}$$

The exhaustive computation therefore evaluates

$$L(\pi)$$

for every one of the twenty-four isotropic nucleotide orders.

The search is finite and complete. No sampling, optimization heuristic, or visual selection is required.

9.2 The Maximum Loop Count

The largest synonymous self-loop count observed among all isotropic orders is

$$\boxed{L_{\max} = 44.}$$

Equivalently,

$$\boxed{\max_{\pi \in S_4} L(\pi) = 44.}$$

This means that no isotropic ordering of the nucleotide alphabet produces more than forty-four nearest-neighbor synonymous edges.

Because every cube contains 144 edges, the maximum synonymous fraction is

$$\frac{44}{144}.$$

Thus,

$$\boxed{f_{\max} = \frac{11}{36} \approx 0.3056.}$$

Approximately 30.6% of the nearest-neighbor edges in the maximizing cube preserve the translated label.

The remaining edges are nonsynonymous:

$$144 - 44 = 100.$$

Therefore, the maximizing multigraph contains

$$44$$

self-loop edges and

$$100$$

edges connecting different translated labels.

The value 44 is not a theoretical upper bound derived from lattice size alone. It is the observed maximum produced by the standard genetic-code assignments under exhaustive isotropic enumeration.

A different mapping of codons to amino acids would generally produce a different loop-count distribution.

9.3 The Maximizing Orders

Exactly two oriented nucleotide orders attain the maximum:

$$\boxed{G - A - C - U}$$

and

$$\boxed{U - C - A - G.}$$

In compact notation,

$$\boxed{L(GACU) = 44.}$$

Likewise,

$$\boxed{L(UCAG) = 44.}$$

The complete maximizing set is therefore

$$\boxed{\arg\max_{\pi \in S_4} L(\pi) = \{GACU, UCAG\}.}$$

No other isotropic order reaches forty-four loops.

The result identifies a specific nucleotide path:

$$G \leftrightarrow A,$$
$$A \leftrightarrow C,$$

and

$$C \leftrightarrow U.$$

The reversed order contains the same undirected adjacent pairs:

$$U \leftrightarrow C,$$
$$C \leftrightarrow A,$$

and

$$A \leftrightarrow G.$$

Thus, the two maximizing orders define the same undirected nucleotide path viewed in opposite directions.

The maximum is therefore associated with one adjacency structure rather than two unrelated solutions.

9.4 Uniqueness up to Reversal

Define the reversal operator

$$R(\pi_0, \pi_1, \pi_2, \pi_3) = (\pi_3, \pi_2, \pi_1, \pi_0).$$

For the maximizing order,

$$R(G, A, C, U) = (U, C, A, G).$$

Reversal changes the orientation of each lattice axis but preserves all undirected nearest-neighbor relationships.

Therefore,

$$E_\pi = E_{R\pi}$$

as undirected edge sets.

It follows that

$$\boxed{L(\pi) = L(R\pi).}$$

Every isotropic order belongs to a reversal pair:

$$[\pi] = \{\pi, R\pi\}.$$

The twenty-four oriented permutations therefore form twelve undirected equivalence classes.

Within this reversal-reduced search space, the G-A-C-U path is the unique maximum.

This may be written as

$$\boxed{\arg\max_{[\pi] \in S_4/\langle R \rangle} L([\pi]) = [GACU].}$$

The equivalence class is

$$[GACU] = \{GACU, UCAG\}.$$

The principal result can therefore be stated in theorem form:

$$\boxed{L(\pi) \leq 44}$$

for every isotropic order π, with equality only when

$$\pi = GACU$$

or

$$\pi = UCAG.$$

The G-A-C-U path is consequently the unique isotropic maximizer up to mirror reversal.

9.5 Comparison with the Next-Highest Orders

The loop-count distribution shows that the maximum is not part of a broad tie.

The two maximizing orders produce

$$44$$

loops.

The next-highest orders produce

$$42$$

loops:

$$\boxed{L(CUAG) = 42,}$$

and

$$\boxed{L(GAUC) = 42.}$$

These two orders are reversals of one another.

The difference between the maximum and the next-highest value is

$$\boxed{44 - 42 = 2.}$$

Several additional orders produce forty-one loops:

$$\boxed{L(AGCU) = 41,}$$
$$\boxed{L(AGUC) = 41,}$$
$$\boxed{L(CUGA) = 41,}$$

and

$$\boxed{L(UCGA) = 41.}$$

These form two reversal pairs:

$$AGCU \leftrightarrow UCGA,$$

and

$$AGUC \leftrightarrow CUGA.$$

The leading portion of the ranking is therefore:

Order	Self-loops
$GACU$	44
$UCAG$	44
$CUAG$	42
$GAUC$	42
$AGCU$	41
$AGUC$	41
$CUGA$	41
$UCGA$	41

The reversal symmetry is visible throughout the ranking. Every order and its reverse have identical loop counts.

The G-A-C-U result is therefore distinguished in two ways:

it attains the largest count,

and

it is the only reversal class attaining that count.

The difference of two loops from the nearest competitor is modest in absolute terms, but exact under the specified statistic. The result should therefore be described as a unique isotropic maximum rather than as an overwhelmingly separated maximum.

9.6 Reproducibility

The enumeration can be reproduced with a short Python program using the standard RNA genetic code.

The required inputs are:

the four nucleotides,
the sixty-four codons,
the standard codon-to-label assignments,

and

the twenty-four nucleotide permutations.

For each permutation, the program generates the lattice coordinates and counts each undirected nearest-neighbor edge once.

The implementation should verify several fixed quantities before accepting the loop-count result.

First,

$$\boxed{|\ \mathcal{C}\ | = 64.}$$

Second,

$$\boxed{|\ E_{\pi}\ | = 144.}$$

Third, all twenty-four isotropic permutations must be evaluated:

$$\boxed{|\ S_4\ |= 24.}$$

Fourth, every order must have the same loop count as its reversal:

$$\boxed{L(\pi) = L(R\pi).}$$

Finally, the program should assert

$$\boxed{\max_{\pi \in S_4} L(\pi) = 44.}$$

It should also verify the maximizing set:

$$\boxed{\{\pi : L(\pi) = 44\} = \{GACU, UCAG\}.}$$

For full transparency, the reproducible output should include:

- the loop count for all twenty-four orders;
- the loop count separated by codon axis;
- the complete list of the forty-four G-A-C-U synonymous edges;
- the loop contribution of each amino-acid or stop vertex;
- the source code used for enumeration.

The forty-four-loop result was independently reconstructed in Python from the standard genetic code rather than copied from the original graph output.

The computation also confirmed that the G-A-C-U loops are distributed unevenly across the three codon-position axes. Most arise from third-position synonymous degeneracy, while smaller contributions occur along the first and second axes.

The exhaustive search establishes the graph-theoretic result:

$$\boxed{GACU \text{ is the unique isotropic maximum up to reversal.}}$$

The next section examines how the forty-four loops are distributed among amino acids and what the maximum reveals about synonymous codon compactness.

10. Biological and Topological Interpretation of the Forty-Four Loops

10.1 Distribution of Loops Among Translated Labels

The total of forty-four self-loops is distributed unevenly among the twenty amino-acid vertices and stop X. This distribution reflects the degeneracy pattern of the standard genetic code together with the particular nucleotide adjacencies created by the G-A-C-U path.

Let

$$\ell_r$$

denote the number of self-loops at translated vertex r.

The largest contributions come from leucine and arginine:

$$\ell_L = 6,$$
$$\ell_R = 6.$$

Serine contributes four loops:

$$\ell_S = 4.$$

Five amino acids contribute three loops each:

$$\ell_A = \ell_G = \ell_P = \ell_T = \ell_V = 3.$$

Isoleucine and stop Xcontribute two loops each:

$$\ell_I = \ell_X = 2.$$

Nine amino acids contribute one loop each:

$$\ell_C = \ell_D = \ell_E = \ell_F = \ell_H = 1,$$
$$\ell_K = \ell_N = \ell_Q = \ell_Y = 1.$$

Methionine and tryptophan contribute no loops:

$$\ell_M = \ell_W = 0.$$

The total is therefore

$$6 + 6 + 4 + 5(3) + 2(2) + 9(1) = 44.$$

Thus,

$$\boxed{\sum_{r \in \mathcal{R}} \ell_r = 44.}$$

The absence of loops at methionine and tryptophan is expected because each is encoded by only one codon. A self-loop requires at least two codons with the same translated label.

The five amino acids contributing three loops each are encoded by complete four-codon families:

$$GCN \to A,$$
$$GGN \to G,$$
$$CCN \to P,$$
$$ACN \to T,$$

and

$$GUN \to V.$$

Because all four third-position nucleotide states encode the same amino acid, they form a four-vertex path along the third codon axis. Any linear order of four nucleotide states contains three consecutive intervals. These families therefore contribute three third-position loops regardless of the particular nucleotide permutation.

Leucine, arginine, and serine also contain four-codon subfamilies and therefore share this three-loop foundation. Their additional codons create further order-dependent loops.

The forty-four-loop total consequently combines two types of contribution:

loops forced by complete synonymous blocks

and

additional loops created by the G-A-C-U adjacency path.

This distinction explains why all nucleotide orders produce some loops, while only G-A-C-U and U-C-A-G attain the maximum.

10.2 Third-Position Degeneracy

Most of the forty-four self-loops occur along the third codon-position axis.

Let

$$L_1, L_2, L_3$$

denote the numbers of synonymous edges parallel to the first, second, and third codon axes.

For G-A-C-U,

$$\boxed{L_1 = 4,}$$
$$\boxed{L_2 = 1,}$$

and

$$\boxed{L_3 = 39.}$$

Therefore,

$$4 + 1 + 39 = 44.$$

The proportion of loops parallel to the third axis is

$$\frac{39}{44} \approx 0.886.$$

Thus, approximately 88.6% of the G-A-C-U self-loops arise from third-position substitutions.

This concentration reflects the well-known degeneracy of the third codon position. Many amino acids are encoded by codons that share the first two nucleotides and differ only at the third position.

The complete four-codon families each form a third-position line:

$$NNG \leftrightarrow NNA \leftrightarrow NNC \leftrightarrow NNU$$

under the G-A-C-U order.

When all four codons on such a line encode the same amino acid, the line contributes three synonymous edges.

The G-A-C-U path also places the two common third-position synonym pairs next to one another:

$$G \leftrightarrow A,$$

and

$$C \leftrightarrow U.$$

These are the two purine and pyrimidine pairs frequently appearing in twofold-degenerate codon families.

For example, phenylalanine is encoded by

$$UUU$$

and

$$UUC.$$

Because $C$and $U$are adjacent in the G-A-C-U path, this synonymous pair contributes one loop.

Glutamate is encoded by

$$GAA$$

and

$$GAG.$$

Because $G$and $A$are adjacent, this pair also contributes one loop.

The same relationship applies to the nine two-codon amino-acid families contributing one loop each:

$$C,\ D,\ E,\ F,\ H,\ K,\ N,\ Q,\ Y.$$

The G-A-C-U path is therefore especially favorable to third-position degeneracy because it contains both major synonymous pairs as direct adjacencies:

$$\boxed{G - A}$$

and

$$\boxed{C - U.}$$

Isoleucine is encoded by three codons:

$$AUU, AUC, AUA.$$

Their third-position states are

$$U,\ C,\ A.$$

Under G-A-C-U, these occur consecutively as

$$A - C - U$$

when read along the relevant segment of the undirected path. The three isoleucine codons therefore contain two adjacent synonymous pairs:

$$AUA \leftrightarrow AUC,$$

and

$$AUC \leftrightarrow AUU.$$

Hence,

$$\ell_I = 2.$$

The dominance of the third axis is therefore not merely a general consequence of wobble. The exact number of retained synonymous edges depends on whether the third-position nucleotide pairs used by the standard code are adjacent in the selected axis order.

10.3 Sixfold-Degenerate Amino Acids

Leucine, arginine, and serine are each encoded by six codons, but their six codons are divided into more than one codon block.

This makes them especially informative for distinguishing degeneracy from topological arrangement.

Leucine is encoded by the four-codon family

$$CUN$$

and the two-codon family

$$UUA,\ UUG.$$

The CUNblock contributes three third-position loops.

The two-codon branch contributes one additional third-position loop because $A$and G are adjacent:

$$UUA \leftrightarrow UUG.$$

Two first-position bridges also occur:

$$CUA \leftrightarrow UUA,$$

and

$$CUG \leftrightarrow UUG.$$

These bridges are present because $C$and $U$are adjacent in the G-A-C-U path.

Therefore,

$$\ell_L = 3 + 1 + 2.$$

Hence,

$$\boxed{\ell_L = 6.}$$

Arginine is encoded by the four-codon family

$$CGN$$

and the two-codon family

$$AGA,\ AGG.$$

The four-codon block contributes three third-position loops.

The two-codon branch contributes one additional loop:

$$AGA \leftrightarrow AGG.$$

Two first-position bridges connect the blocks:

$$AGA \leftrightarrow CGA,$$

and

$$AGG \leftrightarrow CGG.$$

These bridges are present because $A$and $C$are adjacent in the G-A-C-U path.

Therefore,

$$\ell_R = 3 + 1 + 2.$$

Hence,

$$\boxed{\ell_R = 6.}$$

Serine is encoded by the four-codon family

$$UCN$$

and the two-codon family

$$AGU,\ AGC.$$

The UCN block contributes three third-position loops.

The two-codon branch contributes one loop because $C$and $U$are adjacent:

$$AGC \leftrightarrow AGU.$$

The two serine blocks are not joined by a single ordered-cube edge because corresponding codons in the two blocks differ at more than one codon position.

Therefore,

$$\ell_S = 3 + 1.$$

Hence,

$$\boxed{\ell_S = 4.}$$

The sixfold-degenerate amino acids demonstrate why codon count alone does not determine loop count.

All three have six codons, yet

$$\ell_L = 6,$$
$$\ell_R = 6,$$

and

$$\ell_S = 4.$$

Leucine and arginine attain larger loop counts because the G-A-C-U path creates single-position bridges between their four-codon and two-codon blocks.

Serine lacks comparable bridges.

The topology therefore detects organization that is not visible from degeneracy numbers alone.

10.4 Synonymous Compactness

The self-loop count measures local synonymous compactness within the ordered cube.

For translated label r, define its codon set by

$$\mathcal{C}_r = \{c : \phi(c) = r\}.$$

Its degeneracy is

$$n_r = | \mathcal{C}_r |.$$

The loop count

$$\ell_r(\pi)$$

measures how many nearest-neighbor edges of the ordered cube have both endpoints in $\mathcal{C}_r$.

Thus,

$$\ell_r(\pi) = \text{number of local synonymous connections for } r.$$

The total compactness statistic is

$$L(\pi) = \sum_r \ell_r(\pi).$$

A high value of $L(\pi)$ means that many codons with identical translated output occupy neighboring lattice sites.

The statistic is local because it counts only nearest-neighbor relationships. Two synonymous codons may be close in another sense but contribute nothing unless they are directly joined by one ordered-cube edge.

The G-A-C-U result states that this local synonymous compactness is maximal among isotropic nucleotide paths:

$$\boxed{L(GACU) = 44.}$$

The mirror path gives the same value:

$$\boxed{L(UCAG) = 44.}$$

For comparison, the minimum isotropic loop count is

25.

Thus, the isotropic orders span the range

$$\boxed{25 \leq L(\pi) \leq 44.}$$

The difference between the minimum and maximum is

$$44 - 25 = 19.$$

This range shows that nucleotide ordering has a substantial effect on how many synonymous pairs are represented as local neighbors.

A foundation of twenty-five loops persists even in the least favorable isotropic orders. This baseline arises from codon blocks whose degeneracy guarantees some adjacency under any linear nucleotide path.

The G-A-C-U path adds nineteen loops beyond this minimum by placing the relevant synonymous nucleotide pairs and inter-block bridges next to one another.

The compactness statistic therefore contains both invariant and order-sensitive components:

$$\boxed{L(\pi) = L_{\text{baseline}} + L_{\text{order}}(\pi).}$$

For the isotropic search,

$$L_{\text{baseline}} = 25.$$

For G-A-C-U,

$$L_{\text{order}}(GACU) = 19.$$

Thus,

$$25 + 19 = 44.$$

This decomposition helps explain why the leading loop counts differ by only a few edges even though the complete range is much broader. Many loops are fixed by the block structure of the code, while the ranking is determined by the smaller order-sensitive component.

10.5 Relationship to Codon Degeneracy

Degeneracy and synonymous compactness are related but not identical.

Degeneracy is the number of codons assigned to a translated label:

$$n_r = | \mathcal{C}_r |.$$

Compactness is the number of nearest-neighbor connections among those codons:

$$\ell_r(\pi).$$

For any translated label,

$$0 \leq \ell_r(\pi).$$

A loose upper bound is

$$\ell_r(\pi) \leq \binom{n_r}{2}.$$

However, most synonymous codon pairs cannot be nearest neighbors because they differ at more than one position or because their differing nucleotides are not adjacent under π.

A one-codon assignment necessarily has

$$\ell_r = 0.$$

A two-codon assignment may have

$$\ell_r = 1$$

if its two codons are adjacent, or

$$\ell_r = 0$$

if they are not.

A complete four-codon family differing only at one position has

$$\ell_r = 3$$

under every linear order because four states arranged on a path always contain three edges.

A six-codon family may have substantially different compactness depending on whether its codon blocks are connected.

The distinction is illustrated by the three sixfold-degenerate amino acids:

$$n_L = n_R = n_S = 6,$$

but

$$\ell_L = \ell_R = 6,$$

whereas

$$\ell_S = 4.$$

The loop statistic therefore measures the arrangement of degeneracy, not merely its amount.

This is the main biological meaning of the collapsed multigraph. The conventional codon table reports how many codons encode each amino acid. The multigraph additionally reports how those codons are locally connected under a specified nucleotide geometry.

10.6 Limits of the Loop Statistic

The self-loop count is exact and reproducible under its stated definition, but it measures only one aspect of genetic-code topology.

It counts exact preservation of translated identity. It does not assign partial credit when adjacent codons encode different but chemically similar amino acids.

A substitution from one hydrophobic residue to another is counted in the same nonsynonymous category as a substitution from a hydrophobic residue to a charged residue.

A complementary physicochemical edge statistic could assign a cost based on property difference. For property vector $\mathbf{q}_r$, define

$$C_Q(\pi) = \sum_{e \in E_\pi} \Delta_Q\,(e),$$

where $\Delta_Q(e)$measures the physicochemical difference between the translated endpoints.

For example,

$$\Delta_Q(e) = \| \mathbf{q}_{\phi(c)} - \mathbf{q}_{\phi(d)} \|^2.$$

The loop count also treats every cube edge equally. It does not distinguish transitions from transversions, incorporate mutation frequencies, or weight the three codon positions differently.

It does not use codon-usage frequencies. A synonymous edge involving rare codons contributes the same amount as one involving common codons.

The statistic includes stop X as a translated label. The two stop loops therefore contribute to the total of forty-four. Analyses excluding stop should be reported separately rather than silently changing the definition.

The result is also limited to isotropic nucleotide orders. The same order is imposed on all three codon axes.

Allowing independent orders on the three axes expands the search space to

$$(4!)^3 = 13\,824.$$

The G-A-C-U result should therefore be stated precisely:

G-A-C-U is the unique maximum up to reversal among isotropic orders.

The loop count does not establish optimality under every graph criterion. Other possible measures include:

mean distance among synonymous codons,
number of connected synonymous components,
total connector length,
physicochemical smoothness,

and

weighted mutation cost.

The strength of the loop statistic lies in its simplicity. It requires no adjustable parameters, biochemical weights, or fitted thresholds.

Each edge either preserves the translated label or it does not.

The G-A-C-U maximum is therefore a clear and exact topological statement:

$$\boxed{\max_{\pi \in S_4} L(\pi) = 44.}$$

Its interpretation should remain equally precise. It demonstrates maximum local synonymous adjacency within the defined isotropic ordered-cube model.

11. Controls, Alternative Graph Measures, and Extensions

11.1 Separation of Universal and Code-Specific Results

The analysis combines mathematical properties of complete triplet space with biological properties of the standard genetic code. These two levels must be separated clearly.

The complete codon matrix

$$D \in \{0,1\}^{64\times 12}$$

would be unchanged if the amino-acid assignments were replaced by a different genetic code.

Its rank,

$$\operatorname{rank}(D) = 10,$$

its exact pseudoinverse,

$$D^{+} = \frac{1}{16} D^{T} - \frac{1}{96} J,$$

and its Hamming-distance projection kernel,

$$H_{cd} = \frac{5}{32} - \frac{d_H(c,d)}{16},$$

are properties of complete four-letter triplet space.

They do not depend on which amino acid is assigned to any particular codon.

The biologically specific results enter through the translation map

$$\phi: \mathcal{C} \to \mathcal{R}.$$

Equivalently, they enter through the codon-to-amino-acid assignment matrix G.

The property field

$$Y = GQ$$

depends on the actual genetic code. Therefore, the nucleotide-position coefficient matrix

$$B = D^{+}GQ$$

also depends on the biological assignments.

The strong second-position hydropathy effect is code specific. The value

$$\beta_{2,U} \approx 4.02$$

arises because the standard code places several strongly hydrophobic amino acids in the NUN family.

Similarly, the graph-theoretic maximum

$$L(GACU) = 44$$

depends on the placement of synonymous codons within the standard code.

The empty $4 \times 4 \times 4$lattice has no preferred nucleotide path. The preference for G-A-C-U appears only after the codons are labeled by translated identity.

The results should therefore be classified as follows:

$$\boxed{\text{rank, pseudoinverse, and Hamming kernel} = \text{universal word-space mathematics}}$$

and

$$\boxed{\text{property mapping and loop maximum} = \text{standard-code organization.}}$$

This distinction prevents a purely combinatorial identity from being mistaken for a biological finding. The substantive biological question is how strongly the actual assignment map G aligns

amino-acid chemistry and synonymous degeneracy with the mathematical structure supplied by D.

11.2 Randomized-Code Controls

The observed physicochemical and topological patterns should be compared with randomized genetic codes.

A randomized code replaces the standard assignment map ϕ with an alternative map

$$\phi^{(k)}.$$

The corresponding assignment matrix is

$$G^{(k)}.$$

For amino-acid property matrix Q, the randomized codon property field is

$$Y^{(k)} = G^{(k)}Q.$$

Its additive nucleotide-position reconstruction is

$$\hat{Y}^{(k)} = DD^{+}G^{(k)}Q.$$

A simple physicochemical score is the fraction of property variation captured by the additive projection.

For one centered property vector $\mathbf{y}^{(k)}$, define

$$R_k^2 = 1 - \frac{\| \mathbf{y}^{(k)} - DD^{+}\mathbf{y}^{(k)} \|^2}{\| \mathbf{y}^{(k)} \|^2}.$$

The standard-code value can then be compared with the randomized distribution.

An empirical upper-tail probability is

$$\hat{p} = \frac{1 + \sum_{k=1}^{N} I(R_k^2 \geq R_{\text{obs}}^2)}{N + 1}.$$

Here, $I(\cdot)$ is an indicator equal to one when the stated condition is true.

Several randomization levels should be considered.

An unrestricted randomization reassigns translated labels among the sixty-four codons without preserving degeneracy. This provides a broad baseline but generates many biologically unrealistic codes.

A degeneracy-preserving randomization retains the number of codons assigned to each amino acid and stop. If amino acid $r$has degeneracy n_r, every randomized code retains the same n_r.

A synonymous-block-preserving randomization treats complete codon blocks as units. This preserves more of the third-position structure while changing which amino acids occupy those blocks.

A stop-preserving randomization leaves the three termination codons fixed while permuting amino-acid assignments among the sixty-one sense codons.

Each null model answers a different question. The strongest interpretation would come from a result that remains unusual under progressively more constrained randomizations.

The graph statistic should be tested in the same manner. For randomized code k, define

$$L^{(k)}(\pi) = \sum_{e \in E_\pi} I_{\phi^{(k)}}(e).$$

The best isotropic loop count for that randomized code is

$$L_{\max} = \max_{\pi \in S_4} L^{(k)}(\pi).$$

The observed value is

$$L_{\max} = 44.$$

The randomized comparison should allow every code to select its own best nucleotide order. This avoids giving the standard code an unfair advantage by optimizing only the observed code.

The corresponding probability is

$$\hat{p}_L = \frac{1 + \sum_{k=1}^{N} I(L_{\max} \geq 44)}{N + 1}.$$

This test asks whether the standard code achieves unusually high maximum synonymous compactness relative to randomized codes with comparable degeneracy structure.

11.3 Alternative Physicochemical Scales

The original analysis used the Kyte–Doolittle hydropathy scale. This choice should remain central because it reproduces the historical result.

However, hydropathy is not represented by one unique numerical scale. Alternative scales may be derived from solvent partitioning, membrane-transfer energies, protein structures, residue burial, or statistical frequencies.

Let

$$\mathbf{h}^{(1)}, \mathbf{h}^{(2)}, \ldots, \mathbf{h}^{(m)}$$

represent $m$hydropathy scales.

For each scale, calculate

$$\boldsymbol{\beta}^{(j)} = D^{+} G \mathbf{h}^{(j)}.$$

The principal stability question is whether second-position uracil remains the largest positive contribution:

$$\beta_{2,U}^{(j)} = \max_{p,s} \beta_{p,s}^{(j)}.$$

A simple robustness score is

$$P_U = \frac{1}{m} \sum_{j=1}^{m} I \left[\beta_{2,U}^{(j)} = \max_{p,s} \beta_{p,s}^{(j)} \right].$$

Residue-volume values also vary among published scales. Some represent side-chain volume, some residue volume within proteins, and others van der Waals or partial specific volume.

The numerical value assigned to stop X changes with the selected volume scale because

$$V_X = \frac{1}{20} \sum_{r=1}^{20} V_r.$$

Every analysis should therefore report the exact property table used.

When several properties are combined, they should be standardized so that one property does not dominate merely because of its units.

For property j,

$$z_{r,j} = \frac{q_{r,j} - \mu_j}{\sigma_j}.$$

The qualitative patterns expected to remain stable are:

$$NUN \longrightarrow \text{hydrophobic region},$$
$$NAN \longrightarrow \text{hydrophilic region},$$

and

$$NCN \longrightarrow \text{small-residue region.}$$

Robustness across property scales would show that these patterns are not artifacts of one numerical convention.

11.4 Alternative Topological Measures

The self-loop count measures exact preservation of translated identity. It is transparent and parameter free, but it captures only one aspect of graph organization.

A broader analysis can examine the distances among synonymous codons.

For amino acid r, let

$$\mathcal{C}_r = \{c : \phi(c) = r\}.$$

The mean pairwise cube distance within that synonymous set can be defined as

$$\bar{d}_r(\pi) = \frac{1}{\binom{n_r}{2}} \sum_{\{c,d\} \subset \mathcal{C}_r} d_\pi(c, d).$$

A lower value indicates greater global compactness, even when synonymous codons are not immediate neighbors.

The number of connected components within each synonymous subgraph is another useful statistic.

Let

$$K_r(\pi)$$

be the number of connected components formed by the codons encoding amino acid r.

A value of

$$K_r(\pi) = 1$$

means that all codons for $r$form one connected set under the selected cube topology.

A connector-length statistic can measure the minimum number of lattice edges required to join all synonymous codons.

Physicochemical smoothness can also be evaluated. Let

$$\mathbf{q}_r$$

be a physicochemical property vector for amino acid r.

For edge

$$e = \{c, d\},$$

define the property cost

$$\Delta_Q(e) = \| \mathbf{q}_{\phi(c)} - \mathbf{q}_{\phi(d)} \|^2.$$

The total cube cost is

$$C_Q(\pi) = \sum_{e \in E_\pi} \Delta_Q(e).$$

A low value indicates that neighboring codons tend to encode physicochemically similar amino acids.

The loop statistic rewards only exact identity:

$$I_{\phi}(e) = 1$$

when the two translated labels match.

The smoothness statistic also recognizes conservative nonsynonymous changes.

A combined analysis could therefore report:

$$L(\pi)$$

for exact synonymous preservation and

$$C_{Q}(\pi)$$

for physicochemical continuity.

The G-A-C-U path may be tested under each criterion rather than assuming that one maximum implies optimality under every measure.

11.5 Anisotropic Axis Orderings

The present maximum applies to isotropic cubes, in which all three codon axes use the same nucleotide order.

An anisotropic cube allows separate orders:

$$\pi_1, \pi_2, \pi_3.$$

The number of oriented axis triples is

$$(4!)^3.$$

Therefore,

$$\boxed{(4!)^3 = 13\,824.}$$

For one axis triple, define the edge set

$$E_{\pi_1,\pi_2,\pi_3}.$$

The corresponding loop count is

$$L(\pi_1, \pi_2, \pi_3) = \sum_{e \in E_{\pi_1, \pi_2, \pi_3}} I_\phi\,(e).$$

The unrestricted axis-order maximum is

$$L_{\max} = \max_{\pi_1, \pi_2, \pi_3 \in S_4} L(\pi_1, \pi_2, \pi_3).$$

This extension is important because the three codon positions have different biological roles.

The second position organizes much of amino-acid hydropathy.

The third position carries most synonymous degeneracy.

The first position frequently distinguishes amino acids within broad second-position classes.

The nucleotide order best suited to one codon position may therefore differ from the best order at another.

An anisotropic maximum larger than forty-four would not invalidate the isotropic result. It would refine its scope:

$GACU$ is the isotropic maximum.

If the global anisotropic maximum also used G-A-C-U on all three axes, the topological result would become stronger.

The search of $13\,824$ cases is computationally small and can be performed exhaustively.

The output should include the maximizing axis triples, their loop counts, and the separate contributions

$$L_1, L_2, L_3.$$

This would reveal whether different axis orders are favored by first-, second-, and third-position codon structure.

11.6 Statistical Reporting

The final analysis should report complete effect sizes rather than only whether one statistic is maximal.

For the property mapping, report:

$$r,$$
$$R^2,$$

the residual norm,

$$\| \mathbf{r} \|,$$

and the nucleotide-position coefficient matrix.

For randomized-code comparisons, report the observed statistic, randomized mean, randomized standard deviation, empirical probability, and standardized effect size.

A standardized score can be written as

$$Z = \frac{T_{\mathrm{obs}} - \mu_{\mathrm{null}}}{\sigma_{\mathrm{null}}}.$$

For the loop analysis, report the entire distribution of

$$L(\pi)$$

across all twenty-four isotropic orders, not only the maximum.

The result should include:

$$L_{\max} = 44,$$
$$L_{\mathrm{next}} = 42,$$

and

$$L_{\min} = 25.$$

The full range is

$$44 - 25 = 19.$$

The gap between the maximum and the next-highest value is

$$44 - 42 = 2.$$

These quantities describe both the uniqueness and the degree of separation of the G-A-C-U solution.

All computational analyses should specify:

- the exact genetic-code table;
- nucleotide ordering conventions;
- whether RNA U or DNA T notation is used;
- whether stop X is included;
- hydropathy and residue-volume scales;
- randomization constraints;
- number of randomized codes;
- random-number seed;
- treatment of reversal symmetry;
- complete source code.

The central results are simple enough to be verified independently:

$$\begin{aligned}
&\operatorname{rank}(D) = 10, \\
&D^{+} = \frac{1}{16} D^{T} - \frac{1}{96} J, \\
&|\, E_{\pi} \,| = 144,
\end{aligned}$$

and

$$\max_{\pi \in S_4} L(\pi) = 44.$$

The purpose of expanded controls is not to replace these exact findings. It is to determine how unusual the biologically specific alignments are relative to explicitly defined alternatives.

12. Synthesis and Conclusions

12.1 Nucleotide-Determined Amino-Acid Chemistry

The first major result of this analysis is that amino-acid physicochemical properties are substantially recoverable from codon nucleotide identities and positions.

The original hydropathy analysis represented each of the sixty-one sense codons through twelve binary variables: four possible nucleotides at each of three codon positions. The nucleotide-position coefficients were obtained from

$$\boldsymbol{\beta}_{61} = D_{61}^{+}\mathbf{h}_{61}.$$

The strongest positive coefficient was associated with uracil in the second codon position:

$$\boxed{\beta_{2,U} \approx 4.02.}$$

The strongest negative second-position coefficient was associated with adenine:

$$\boxed{\beta_{2,A} \approx -2.92.}$$

These coefficients quantify a broad organization already visible in the genetic-code table. Codons of the form NUN predominantly encode hydrophobic amino acids, whereas codons of the form NAN predominantly encode hydrophilic amino acids.

For codon

$$c = (c_1, c_2, c_3),$$

the additive hydropathy prediction is

$$\boxed{\widehat{H}(c) = \beta_{1,c_1} + \beta_{2,c_2} + \beta_{3,c_3}.}$$

The model does not reconstruct every amino-acid property exactly. Its importance lies in the amount of chemically interpretable structure captured by a low-dimensional additive representation.

The paired hydropathy and residue-volume analysis extends the result into two-dimensional physicochemical space. The four second-position codon families occupy distinguishable regions:

$NUN \longrightarrow$ hydrophobic region,
$NCN \longrightarrow$ small-residue region,
$NAN \longrightarrow$ hydrophilic region,

and

$NGN \longrightarrow$ more heterogeneous region.

Thus, the second codon nucleotide functions as a coarse physicochemical classifier. It does not uniquely determine amino-acid identity, but it strongly restricts the hydropathy and volume range of the encoded residue.

The central biological mapping is

$$\boxed{B = D^{+}GQ.}$$

Here, Q contains amino-acid properties, G assigns those properties to codons, and D^{+}maps the resulting codon property fields onto nucleotide-position coefficients.

The complete transformation is

$$\boxed{Q \overset{G}{\rightarrow} Y \overset{D^{+}}{\rightarrow} B \overset{D}{\rightarrow} \hat{Y}.}$$

This equation summarizes the principal physicochemical result of the paper: amino-acid chemistry can be expressed, to a substantial degree, in the coordinate system of codon nucleotides and positions.

12.2 Low-Dimensional Codon Algebra

The complete codon space contains all sixty-four triplets formed from four nucleotides at three positions. Its one-hot representation is

$$D \in \{0,1\}^{64\times 12}.$$

Although the matrix has twelve columns, two exact dependencies arise because the four nucleotide indicators within each positional block sum to the same all-ones vector.

Therefore,

$$\boxed{\text{rank}\,(D) = 10.}$$

The ten independent dimensions consist of one overall mean and three nucleotide contrasts at each of three codon positions:

$$\boxed{10 = 1 + 3 + 3 + 3.}$$

The complete codon matrix has the exact Moore–Penrose pseudoinverse

$$\boxed{D^{+} = \frac{1}{16} D^{T} - \frac{1}{96} J.}$$

Every pseudoinverse entry takes one of two values:

$$\boxed{(D^{+})_{(p,s),c} = \begin{cases} \frac{5}{96}, & c_p = s, \\ -\frac{1}{96}, & c_p \neq s. \end{cases}}$$

The exact formula is a consequence of complete factorial balance. Each nucleotide occurs sixteen times at each codon position, and every ordered nucleotide pair occurs four times across any two positions.

The historical sixty-one-codon analysis and the complete sixty-four-codon formulation differ slightly in their treatment of stop codons, but this distinction need not dominate the interpretation. Both implement the same fundamental transformation from a codon-level property field to nucleotide-position coefficients.

The complete-cube formulation is especially useful because it exposes the exact geometry of the projection

$$H = DD^{+}.$$

For codons c and d,

$$\boxed{H_{cd} = \frac{5}{32} - \frac{d_H(c,d)}{16}.}$$

Thus, the additive projection coefficient between two codons depends only on their Hamming distance.

The algebraic and geometric views are therefore equivalent:

$$\boxed{\text{exact pseudoinverse} \leftrightarrow \text{Hamming-distance projection.}}$$

The matrix representation is order independent. It does not require the nucleotides to be arranged along a particular path. This distinguishes it from the ordered codon cube developed in the topological analysis.

12.3 Codon-Cube Topology

The sixty-four codons can be placed on a $4 \times 4 \times 4$ lattice whose three axes represent the three codon positions.

An isotropic cube applies the same nucleotide order to all three axes.

For nucleotide order π, the ordered codon graph is

$$G_\pi = (\mathcal{C}, E_\pi).$$

Every isotropic cube contains

$$64$$

codon vertices and

$$\boxed{|\, E_\pi \,| = 144}$$

nearest-neighbor edges.

The edge count follows from

$$\boxed{|\, E_\pi \,| = 3(4-1)4^2.}$$

The ordered cube differs from the full Hamming graph. The Hamming graph connects all codons differing at one position and contains 288edges. The ordered cube retains only substitutions between consecutive nucleotides along the selected axis path.

The topology therefore depends on the nucleotide order.

For the G-A-C-U cube, each axis follows

$$G \leftrightarrow A,$$

$$A \leftrightarrow C,$$

and

$$C \leftrightarrow U.$$

The sixty-four codon vertices are then collapsed according to their translated labels. The resulting multigraph contains twenty amino-acid vertices and stop X:

$$\boxed{| V_{\text{collapsed}} |= 21.}$$

All 144 codon edges are retained. Several codon-level edges may connect the same pair of translated labels, so parallel edges occur naturally.

When adjacent codons encode the same label, their edge becomes a self-loop.

Define the synonymous-edge indicator for edge $e$by

$$\boxed{I_\phi(e) = \mathbf{1}[\text{the endpoints of } e \text{ have the same translated label}].}$$

The total self-loop count is

$$\boxed{L(\pi) = \sum_{e \in E_\pi} I_\phi\,(e).}$$

This statistic measures local synonymous compactness. It counts the number of ordered-cube adjacencies that preserve amino-acid or stop identity.

The multigraph extends the conventional concept of degeneracy. Degeneracy counts how many codons encode one amino acid. The multigraph additionally shows how those codons are connected.

Thus,

$$\boxed{\text{degeneracy} \neq \text{topological compactness.}}$$

Two amino acids may have equal degeneracy but different self-loop counts because their codons occupy different geometric arrangements.

12.4 The G-A-C-U Maximum

The four nucleotides admit

$$4! = 24$$

isotropic linear orders.

Every order was evaluated exhaustively using the same codon assignments, nearest-neighbor rule, and loop-count statistic.

The maximum is

$$\boxed{\max_{\pi \in S_4} L(\pi) = 44.}$$

The only oriented orders attaining this value are

$$\boxed{GACU}$$

and

$$\boxed{UCAG.}$$

Therefore,

$$\boxed{\arg\max_{\pi \in S_4} L(\pi) = \{GACU, UCAG\}.}$$

U-C-A-G is the complete reversal of G-A-C-U. Reversal preserves all undirected nucleotide adjacencies, so the two orders represent one mirror-equivalent topology.

The correct uniqueness statement is therefore

G-A-C-U is the unique isotropic maximum up to reversal.

The next-highest reversal pair produces forty-two loops:

$$L(CUAG) = L(GAUC) = 42.$$

The maximum therefore exceeds the nearest alternative by two loops.

The isotropic loop counts span a substantially broader range:

$$\boxed{25 \leq L(\pi) \leq 44.}$$

The difference between the least and most compact isotropic embeddings is

$$44 - 25 = 19.$$

This range demonstrates that nucleotide ordering has a substantial effect on local synonymous connectivity.

The forty-four G-A-C-U loops are distributed unevenly among the translated labels.

Leucine and arginine contribute six each:

$$\ell_L = \ell_R = 6.$$

Serine contributes four:

$$\ell_S = 4.$$

Alanine, glycine, proline, threonine, and valine contribute three each:

$$\ell_A = \ell_G = \ell_P = \ell_T = \ell_V = 3.$$

Isoleucine and stop Xcontribute two each:

$$\ell_I = \ell_X = 2.$$

Nine two-codon amino-acid families contribute one loop each.

The axis decomposition is

$$\boxed{L_1 = 4,}$$
$$\boxed{L_2 = 1,}$$

and

$$\boxed{L_3 = 39.}$$

Thus, most loops arise from third-position degeneracy:

$$\frac{39}{44} \approx 0.886.$$

The G-A-C-U path is especially effective because it places both major third-position synonymous pairs next to one another:

$$G \leftrightarrow A$$

and

$$C \leftrightarrow U.$$

It also creates first-position bridges between the codon blocks of leucine and arginine, increasing their self-loop contributions beyond those of serine.

The forty-four-loop maximum is therefore not a single unexplained number. It arises from identifiable features of codon degeneracy, third-position pairing, and inter-block connectivity.

12.5 Unified Interpretation

The principal findings form a connected mathematical sequence.

Amino-acid properties are first assigned to codons:

$$\boxed{Y = GQ.}$$

The codon property fields are mapped onto nucleotide-position coefficients:

$$\boxed{B = D^{+}GQ.}$$

The additive component is reconstructed:

$$\boxed{\hat{Y} = DD^{+}GQ.}$$

The projection is governed by Hamming distance:

$$\boxed{H_{cd} = \frac{5}{32} - \frac{d_H(c,d)}{16}.}$$

The codons are then embedded in an ordered cube, collapsed by translated identity, and evaluated through synonymous self-loops:

$$\boxed{L(\pi) = \sum_{e \in E_\pi} I_\phi\,(e).}$$

The resulting isotropic extremum is

$$\boxed{L(GACU) = L(UCAG) = 44.}$$

These results describe different aspects of one object.

The property mapping concerns amino-acid chemistry.

The rank-ten algebra concerns additive nucleotide-position coordinates.

The Hamming kernel concerns order-independent symbolic distance.

The codon cube concerns order-dependent geometric adjacency.

The multigraph concerns the topology of synonymous translation.

Together, they show that the standard genetic code can be studied simultaneously as

a physicochemical map,
a low-rank algebraic system,
a Hamming geometry,

and

an ordered multigraph.

The graph-theoretic result is particularly important because it reveals organization not captured by the coefficient analysis alone. Hydropathy mapping concerns broad amino-acid property classes. The self-loop analysis concerns exact preservation of translated identity.

The two analyses are therefore complementary:

$$\boxed{\text{physicochemical alignment} + \text{synonymous topological compactness.}}$$

The standard code displays both.

12.6 Final Conclusion

The standard genetic code is more mathematically structured than its conventional lookup-table presentation suggests.

Amino-acid hydropathy and residue volume are substantially organized by codon position, especially by the second nucleotide. Uracil at the second position provides the strongest hydrophobic contribution, while adenine provides a strong hydrophilic contribution.

The complete codon matrix has rank ten and an exact Moore–Penrose pseudoinverse. Its projection kernel is determined solely by codon Hamming distance.

When the codons are embedded in isotropic ordered cubes and collapsed to a twenty-one-vertex translated multigraph, G-A-C-U and its reversed mirror U-C-A-G uniquely maximize synonymous self-loops.

The core results are

$$\boxed{\operatorname{rank}(D) = 10,}$$

$$\boxed{D^{+} = \frac{1}{16} D^{T} - \frac{1}{96} J,}$$

$$\boxed{H_{cd} = \frac{5}{32} - \frac{d_H(c,d)}{16},}$$

and

$$\boxed{\max_{\pi \in S_4} L(\pi) = 44.}$$

The maximizing set is

$$\boxed{\{GACU, UCAG\}.}$$

These equations summarize the algebraic, geometric, and topological structure developed throughout the paper.

The most direct conclusion is that the standard genetic code joins nucleotide composition, amino-acid chemistry, symbolic distance, synonymous degeneracy, and graph connectivity within one coherent mathematical framework.

13. References

1. Yang, M. M., Coleman, W. J., and Youvan, D. C. Genetic coding algorithms for engineering membrane proteins. In: Michel-Beyerle, M. E., editor. Reaction Centers of Photosynthetic Bacteria. Springer Series in Biophysics, Volume 6. Berlin and Heidelberg: Springer-Verlag; 1990. pp. 209–218. DOI: 10.1007/978-3-642-61297-8_21.

2. Kyte, J., and Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. Journal of Molecular Biology. 1982;157(1):105–132. DOI: 10.1016/0022-2836(82)90515-0.

3. Zamyatnin, A. A. Protein volume in solution. Progress in Biophysics and Molecular Biology. 1972;24:107–123. DOI: 10.1016/0079-6107(72)90005-3.

4. Zamyatnin, A. A. Amino acid, peptide, and protein volume in solution. Annual Review of Biophysics and Bioengineering. 1984;13:145–165. DOI: 10.1146/annurev.bb.13.060184.001045.

5. Moore, E. H. On the reciprocal of the general algebraic matrix. Bulletin of the American Mathematical Society. 1920;26:394–395.

6. Penrose, R. A generalized inverse for matrices. Proceedings of the Cambridge Philosophical Society. 1955;51:406–413. DOI: 10.1017/S0305004100030401.

7. Golub, G. H., and Reinsch, C. Singular value decomposition and least squares solutions. Numerische Mathematik. 1970;14:403–420. DOI: 10.1007/BF02163027.

8. Hamming, R. W. Error detecting and error correcting codes. Bell System Technical Journal. 1950;29(2):147–160. DOI: 10.1002/j.1538-7305.1950.tb00463.x.

9. Youvan, D. C., and Juvon, C. Decryption and Topology of the Genetic Code. ResearchGate preprint. 2022. DOI: 10.13140/RG.2.2.34261.91361.

10. Youvan, D. C. Graph Analysis of the Genetic Code Using Closed-Loops Shows Isotropic Axes Using the Order GACU Yields Maximum Topological Compactness. ResearchGate preprint. 2022. DOI: 10.13140/RG.2.2.21318.55362.

11. Youvan, D. C. The Genetic Code Has a Formulae-Based Pseudo-Inverse. ResearchGate preprint. 2023. DOI: 10.13140/RG.2.2.18540.03203.

12. Youvan, D. C. Mathematica Code for the Singular Value Decomposition of the Genetic Code for the 20 Kyte–Doolittle Amino Acid Residue Hydropathy Indices to 12 Nucleotide-Determined Values Without Codon Translation: Replicating Our 1990 Publication. ResearchGate preprint. 2023. DOI: 10.13140/RG.2.2.11252.42883.

13. Youvan, D. C. Amino Acids Are Segregated in Hydropathy–Molar Volume Space by the Second Position of the Codon. ResearchGate preprint. 2023. DOI: 10.13140/RG.2.2.10740.94083.

14. Haig, D., and Hurst, L. D. A quantitative measure of error minimization in the genetic code. Journal of Molecular Evolution. 1991;33(5):412–417. DOI: 10.1007/BF02103132.

15. Freeland, S. J., and Hurst, L. D. The genetic code is one in a million. Journal of Molecular Evolution. 1998;47(3):238–248. DOI: 10.1007/PL00006381.

16. Hopp, T. P., and Woods, K. R. Prediction of protein antigenic determinants from amino acid sequences. Proceedings of the National Academy of Sciences of the United States of America. 1981;78(6):3824–3828. DOI: 10.1073/pnas.78.6.3824.

17. Wimley, W. C., and White, S. H. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. Nature Structural Biology. 1996;3:842–848. DOI: 10.1038/nsb1096-842.

18. Harary, F. Graph Theory. Reading, Massachusetts: Addison-Wesley; 1969.

19. Newman, M. E. J. Networks: An Introduction. Oxford: Oxford University Press; 2010. DOI: 10.1093/acprof/9780199206650.001.0001.

20. Hunter, J. D. Matplotlib: A 2D graphics environment. Computing in Science and Engineering. 2007;9(3):90–95. DOI: 10.1109/MCSE.2007.55.

21. Hagberg, A. A., Schult, D. A., and Swart, P. J. Exploring network structure, dynamics, and function using NetworkX. In: Varoquaux, G., Vaught, T., and Millman, J., editors. Proceedings of the Seventh Python in Science Conference. Pasadena, California; 2008. pp. 11–15.

22. Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. Array programming with NumPy. Nature. 2020;585:357–362. DOI: 10.1038/s41586-020-2649-2.

23. National Center for Biotechnology Information. The Genetic Codes. NCBI Taxonomy resources. Bethesda, Maryland: National Library of Medicine.

24. International ImMunoGeneTics Information System. Amino-acid abbreviations, physicochemical characteristics, residue volumes, and hydropathy indices. IMGT Education Resources.

Postscript

Postscript: Mathematical Evidence, Prior Recognition, and Interpretive Bias

The mathematical results presented in this paper are independent of disputes over the origin, meaning, or philosophical implications of the genetic code. The rank of the codon matrix, its exact pseudoinverse, the nucleotide-determined hydropathy coefficients, the Hamming-distance projection, and the exhaustive graph enumeration are derivable or computationally reproducible. Their validity does not depend on the personal beliefs of the investigators who study them.

The interpretation and historical reception of such findings are different matters. Scientific communities do not operate outside cultural, institutional, or ideological pressures. Decisions about which patterns deserve sustained attention, which explanations are considered acceptable, and which implications are regarded as dangerous may be influenced by commitments that are not themselves mathematical.

The physicochemical organization of the genetic code was recognized early. Carl Woese published “Order in the Genetic Code” in 1965, and Mikhail Volkenstein discussed the coding of polar and nonpolar amino acids during the same period. Historical treatments explicitly describe the second-position pattern: codons containing (U) in the middle position tend to encode hydrophobic amino acids, codons containing (A) tend to encode hydrophilic amino acids, codons containing (C) occupy an intermediate region, and codons containing (G) form a more heterogeneous group. The broad relationship between codon position and amino-acid chemistry was therefore visible near the beginning of modern genetic-code research.

These observations are directly continuous with the mathematical analysis developed in this paper. The 1990 singular-value-decomposition calculation quantified the same second-position organization and showed that (U) at the middle codon position is the dominant positive determinant of nucleotide-predicted hydropathy. The later exact pseudoinverse formulation provided an analytical expression for mapping complete codon property fields onto nucleotide-position coefficients. The codon-cube and multigraph analyses then revealed an additional topological organization that was not evident from the conventional table.

The author witnessed a relevant example of institutional and ideological commitment directly. In 1975, while working with Thomas Jukes at the Space Sciences Laboratory in the Berkeley–Richmond area, Jukes stated with pride that he and Francis Crick had “run the Creationists out of California high schools.” This recollection is reported as eyewitness testimony. It is not offered as proof that either scientist suppressed a particular derivation, nor does it establish that they had obtained the exact rank-ten, pseudoinverse, Hamming-kernel, or G-A-C-U results presented

here. It does, however, reveal an explicitly adversarial posture toward one side of a major scientific and educational controversy.

The historical record independently shows that Jukes actively opposed creationist influence in science curricula and textbooks and used prominent scientific forums, including his writing in Nature, to pursue that objective. Crick also participated in later public efforts opposing the inclusion of creation science in public-school science education. Whatever judgment one makes about the merits of that educational campaign, these activities demonstrate that influential geneticists were not detached observers of the surrounding ideological conflict.

This matters because the existence of bias need not take the form of falsifying data or deliberately hiding equations. Bias can operate through selection of problems, preferred terminology, allocation of attention, standards of significance, and limits placed on acceptable interpretation. A strong mathematical order may be acknowledged but described as incidental. A persistent physicochemical correspondence may be treated as secondary to historical contingency. A structural result may remain unexplored because its implications are thought likely to strengthen an opposing cultural or religious position.

The early recognition of order by Woese, Volkenstein, and others makes this concern more specific. The second-position hydropathy pattern was not invisible. It was known in broad form. The question is why such observations did not lead more centrally to a general mathematical treatment of the genetic code as a physicochemical and topological object. The later rank-ten formulation, exact pseudoinverse, Hamming kernel, and graph-theoretic extremum show that the early qualitative observations belonged to a deeper and more unified mathematical structure.

This concern should still be distinguished from an unsupported accusation of deliberate concealment. The available evidence does not prove that Crick, Jukes, or their colleagues had derived the specific mathematics presented in this paper and intentionally withheld it. The stronger and more defensible conclusion is that ideological commitments may have influenced which dimensions of the genetic code were treated as fundamental and which were subordinated to prevailing explanatory frameworks.

The appropriate corrective is not to replace one predetermined interpretation with another. It is to separate mathematical evidence from philosophical preference. The genetic code should be examined through explicit equations, reproducible computation, transparent property scales, complete permutation searches, and clearly defined graph statistics. Interpretations should remain accountable to the measurable structure rather than deciding in advance how much order is permitted to be significant.

The main body of this paper has therefore confined itself to the mathematical and physicochemical results. This postscript records a historical concern: substantial order in the genetic code was recognized early, while the scientific culture surrounding its interpretation was not ideologically neutral. Readers may disagree about the broader implications. The equations, figures, and exhaustive computations remain open to direct inspection.